PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 :	A1	(11) International Publication Number: WO 00/15851
C12Q 1/68, G06F 17/30		(43) International Publication Date: 23 March 2000 (23.03.00)

(21) International Application Number: PCT/US99/21525

(22) International Filing Date: 17 September 1999 (17.09.99)

(30) Priority Data:
60/101,009 17 September 1998 (17.09.98) US
09/398,404 16 September 1999 (16.09.99) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Applications
US 60/101,009 (CIP)
Filed on 17 September 1998 (17.09.98)
US 09/398,404 (CIP)
Filed on 16 September 1999 (16.09.99)

(71) Applicant (*for all designated States except US*): CURAGEN CORPORATION [US/US]; 13th floor, 555 Long Wharf Drive, New Haven, CT 06511 (US).

(72) Inventor; and
(75) Inventor/Applicant (*for US only*): BADER, Joel, S. [US/US]; Apartment 2, 12 Academy Street, New Haven, CT 06511 (US).

(74) Agent: PRINCE, John, T.; Mintz, Levin, Cohn, Ferris, Glovsky and Popeo, P.C., One Financial Center, Boston, MA 02111 (US).

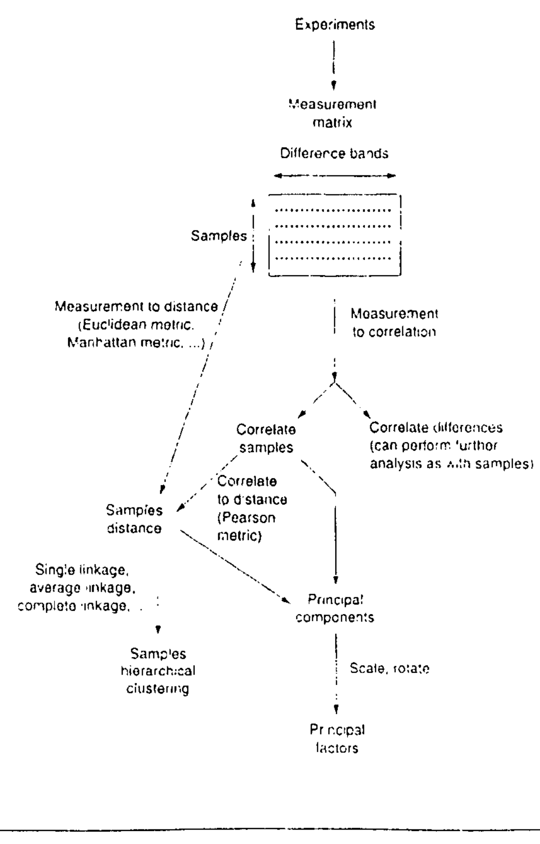
(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: GEOMETRICAL AND HIERARCHICAL CLASSIFICATION BASED ON GENE EXPRESSION

(57) Abstract

The present invention provides a method for generating a representation of the extent of relatedness between at least two classes of cells. The invention also provides a method for generating a representation of the correlation between a first class of cells and a second class of cells. The correlation reflects a change in the nature and amount of nucleic acids present in the classes. In these methods, the cells in each class are chosen from among cells of a given cell type, cells from a given tissue, and cells from a given organ. The methods establish similarities or differences between the classes by defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence, and, in the nucleic acid of each class of cells, determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, as well as a quantitation of the extent to which each fragment is present. The methods then determine the extent of relatedness reflecting the similarities or differences among the classes. The invention further provides display means displaying a representation of the extent of relatedness between the classes of cells, and displaying a representation of the correlation between the first class of cells and the second class of cells. Additionally, the invention provides a representation of the extent of relatedness between the classes of cells, and representation of the correlation between the first class of cells and the second class of cells.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav	TM	Turkmenistan
BF	Burkina Faso	GR	Greece		Republic of Macedonia	TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's	NZ	New Zealand		
CM	Cameroon		Republic of Korea	PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

GEOMETRICAL AND HIERARCHICAL CLASSIFICATION BASED ON GENE EXPRESSION

FIELD OF THE INVENTION

This invention relates to representations of the extent of relatedness between cells, cell lines, tissues, organs, or expressed sequences based on a genomic analysis of gene expression using software algorithm based analysis.

RELATED APPLICATIONS

This application claims priority to both United States Application Serial Number ________, filed September 16, 1999, entitled "GEOMETRICAL AND HIERARCHICAL CLASSIFICATION BASED ON GENE EXPRESSION", and United States Provisional Application Serial Number 60/101,009 filed September 17, 1998, entitled "PHYLOGENOMICS AND PHARMACOGENOMICS", which are incorporated herein by reference in their entirety.

BACKGROUND OF THE INVENTION

The rapid development of genomics and proteomics in recent years has led to a burgeoning of applications making use of the new information provided. A significant area in which such information has been put to use is in the grouping and characterization of pathological states according to the differential expression of genes in such states. A corollary application is in grouping and characterizing the therapeutic effects of known or candidate pharmaceutical agents used in treating various pathologies. Algorithms employing a variety of statistical procedures have been employed to create heuristic displays of the information obtained from such analyses. These displays include large two dimensional, or even higher dimensional arrays in which the elements are coded, for example by false color coding, to represent a particular experimental result. Alternative displays include those in which the experimental data is used to generate cladistic or radiating tree structures as a representation of relatedness. Furthermore, it is also possible to use similar methods to group expressed sequences according to patterns of co-expression over several different biological states.

For example, a system of cluster analysis for genome-wide expression in the yeast *Saccharomyces cerevisiae* and in primary human fibroblasts has been presented by Eisen *et al.* (*Proc. Natl. Acad. Sci. USA* 95:14863-14868 (1998)). In the yeast work, DNA microchip arrays carrying essentially every ORF from this organism were used. Differential expression was studied by varying the physiological state, including the diauxic shift, the mitotic cell division cycle, sporulation, and temperature and reducing shocks. The human fibroblasts were stimulated with serum following serum

starvation, and examined using a microarray with 9,800 cDNAs representing approximately 8,600 distinct human transcripts. Additionally, a further independent variable in these experiments is the time at which an assay point was taken. Data reflecting the differential gene expression in the various studies were analyzed using pairwise average-linkage cluster analysis (Sokal *et al.*, *Univ. Kans. Sci. Bull.* 38:1409-1438 (1958)), which was used to compute a dendrogram that assembles all elements into a single tree.

Colon adenocarcinoma from 40 tumor samples were compared with 22 normal colon tissue samples using Affymetrix DNA chips to which sequences from human cDNAs were bound (Alon *et al.*, *Proc. Natl. Acad. Sci. USA* 96:6745-6750 (June 1999)). 3,200 full-length human cDNAs and 3,400 ESTs are represented in sets of 25-bp fragments, as well as such sequences containing a single base mismatch in the center of the sequence. The gene expression in both the tumor tissue samples and the normal colon samples, was assessed by hybridization. The statistical significance of the correlation between genes was assessed by calculating pairwise correlation coefficients. The clustering of the expressed genes was evaluated using an algorithm based on deterministic-annealing (Rose *et al.*, *Phys. Rev. Lett.* 65:945-948 (1990); Rose, Proc. IEEE 96: 2210-2239 (1998)) to organize the data in a binary tree. Data are presented as a large two-dimensional color coded array, with genes displayed along one dimension and tissue samples along the other; artificial color values are assigned at each array point to indicate the extent of expression in a third dimension. Clustering analysis reveals patterns in the color distribution within the array which is disrupted when various randomization procedures are applied. The clustering of the genes in the data set reveals groups of genes whose expression is correlated across tissue types. The algorithm separated the tissues into distinct clusters.

Pharmacological effects of compounds actually used or being screened for use in cancer chemotherapy were analyzed by cluster analysis at the National Cancer Institute (Weinstein *et al.*, *Science* 275:343-349 (1997)). More than 60,000 compounds were screened against a panel of 60 human cancer cell lines. A 50% growth-inhibitory concentration of a compound in a given cell line, when analyzed across all cell lines, provided detailed information on mechanisms of drug action and drug resistance. Patterns of activity were first analyzed by the COMPARE algorithm (Paull *et al.*, J. Natl. Cancer Inst. 81:1088 (1989); Jayaram, *Biochem. Biophys. Res. Commun.* 186:1600 (1992); Paull *et al.*, In: CANCER CHEMOTHERAPEUTIC AGENTS, Foye (ed.), American Chemical Society, Washington DC, 1993, pp. 1574-1581; Boyd *et al.*, *Drug Dev. Res.* 34:91 (1995)). The procedures developed rely on three databases, an S database characterizing structural information on the candidate compounds, an A database related to the 60 cell lines and a T database including information on molecular targets of action. In an example of the results of the analysis, a three dimensional array displaying compounds

versus targets, with a false color code providing a correlation coefficient in a third dimension for each position in the array, was developed.

Certain problems arise upon consideration of the procedures currently in use for the correlation and clustering of genome-derived attributes. Use of DNA microchips inherently limits any analysis to the sampling of the DNA sequence fragments employed as the capture probes bound to the chips. Detection of any DNA fragment which does not hybridize with one of the capture probes is not possible, so that positive results are potentially lost. Additionally, a mutation or other allelic polymorphism may not bind to the capture probe under conditions of moderate or low stringency, so that again information relating to a positive result may be lost.

For these reasons there is a need for methods of genomic statistical analysis based on more comprehensive accessibility to the genomes of the organisms being studied. Furthermore there remains a need for ways of presenting the information obtained in genomic analyses of relatedness of genes, and in genomic analysis of response to actual or candidate pharmaceutical agents, that includes information gleaned from a comprehensive access to the genomes in question. The present invention addresses these needs, for use is made in the invention of partial and full genomic sequences available from a large number of sequence databases in clustering analysis of the components appearing as independent variables in a particular study.

SUMMARY OF THE INVENTION

The invention provides novel methods of geometric and hierarchical classification between at least two classes of data sets. Data sets may represent cells, nucleic acid sequences, polypeptide sequences, or the like. The invention is able to utilize both standard DNA microchip arrays and non-DNA chip technology to provide input information on nucleic acid moieties of the specified classes of cells. The data are then treated in various ways to provide representations of relatedness that are readily interpretable by the human eye. The invention additionally provides novel methods for generating a representation of the correlation between at least two classes of cells, the correlation reflecting any changes in the composition and amount of nucleic acids present between the classes.

The cell classes may be from different sources for use in comparing differences between various cell populations. These differences include, but are not limited to, species differences, tissue differences, disease state differences, and drug treatment differences. Computer algorithms analyze input data reflecting differences between chosen cell classes and represent them in a meaningful way.

Prior to the present invention, input information was obtained only using DNA-chip technology to analyze the nucleic acids of the cell classes to be compared. Drawbacks to these methods are that identifier sequences need to be already known and isolated, chip technology has size limitations related to the number of the nucleic acids immobilized on the chips, and, once the chips were manufactured, it is virtually impossible to expand nucleic acid parameters. The invention provides the use of GeneCalling™, a non-DNA chip technology, to assay differences between input cell classes. An unexpected result is that GeneCalling™ is able to provide sensitive comparisons between disparate groups above, thereby sidestepping the limitations inherent in the use of DNA chip technology when assaying input nucleic acid population.

The invention provides a novel method for generating the extent of relatedness reflecting similarities or differences in the presence and quantitation of the fragments among the classes by calculating a distance that reflects the amplitude of a difference vector. In a significant embodiment of this method for generating the representation of relatedness, the extent of relatedness is provided by generating a tree structure reflecting the relatedness between any two classes. The branches of the tree structure reflect the difference vectors and are ramified from nodes.

The invention also provides a novel method for generating a representation of the correlation between classes of data sets. In a significant embodiment of the method for generating a representation of the correlation, the correlation is related to a set of orthonormal eigenvectors. In another significant embodiment of the method for generating a representation of the correlation, the representation is a cluster diagram or a dendrogram, and includes a tree structure reflecting the relatedness of the pathways involved in the biochemical or physiological response to a difference between cells of the two classes.

The invention additionally relates to providing geometrical representations of differences between classes of data sets. The geometrical representations encompass, by way of nonlimiting example, principal component analysis and principal factor analysis, as well as reduced dimensional representations derived from them. The geometrical representations are based on differences determined between classes of cells using any method of analyzing for the presence of genes, nucleic acids, or fragments thereof, including nucleic acid microchip arrays and differential display of expressed genes or nucleic acid fragments.

The invention also provides display means for displaying the representation of the extent of relatedness, the correlation, and the geometrical representations of differences between classes of data sets, as well as the representations themselves.

BRIEF DESCRIPTION OF THE DRAWING

Figure 1 is a schematic flow diagram illustrating the principal steps involved in generating the various representations of the invention starting from a set of subsequence-selected fragments found for the samples.

Figure 2 is a schematic flow diagram illustrating the primary steps involved in carrying out a principal component analysis.

Figure 3 illustrates hierarchical clustering of four drugs with sterile water as an outgroup.

Figure 4 is a graphical projection of drug treatments and controls onto principal factors.

DETAILED DESCRIPTION

The present invention relates to methods for preparing representations of the relatedness between cells of any two or more different classes of cells. The classes broadly encompass cells arising in animal and plant organisms, the cells further being normal cells or cells in a diseased state, including tumor cells. They further include cells that have been treated with a putative pharmaceutical agent. The representations are obtained using experimental data that provide size and sequence information on nucleic acid fragments derived from each of the cellular sources. The fragments may be prepared from the nucleic acid content of the cells in each class in any of several ways. For example, in a particularly important embodiment, they may be subjected to digestion by particular pairs of restriction endonucleases; alternatively, in another important embodiment, cell extracts may be subjected to amplification using specially designed primer oligonucleotides. The present invention also relates to methods for preparing representations of the relatedness in terms of co-expression between the nucleic acid fragments so produced.

The invention further relates to the representations provided by these methods, and to display means on which such representations are displayed. The methods for preparing the fragments, such as the use of restriction endonucleases or the application of amplification primers, are chosen to provide subsequence information relating to the ends of the resulting fragments, while size determination provides the length of the fragment. In certain applications of these types of information, the size and subsequence results can optionally be scanned against databases providing known nucleic acid sequences in order to provide the identity of one or more candidate fragments of known complete nucleic acid sequences having the correct length and terminal subsequences (U. S. Patent No. 5,871,697; Shimkets *et al.* 1999 *Nature Biotechnology* 17:798-803). This database look-up step is not a required feature of the current invention. For this reason, the present representations and methods are more comprehensive and

more informative of genomic variations among the samples than those currently known. As described in the Background of the Invention, currently known procedures are restricted in their comprehensiveness to those nucleic acid fragments that are applied to DNA microchips as probe sequences in a given procedure. Except for a narrowly limited set of model organisms with known genome sequence, the number of such probe sequences is considerably fewer than the number of known nucleic acid sequences available in sequence databases and employed in the present invention. Furthermore, even for fully sequenced genomes, genetic variation is not adequately probed with existing DNA microchips. This distinction characterizes an important advantage of the instant invention.

The invention additionally relates to providing geometrical representations of differences between classes of cells. The geometrical representations encompass, by way of nonlimiting example, principal component analysis and principal factor analysis, as well as reduced dimensional representations derived from them. The geometrical representations are based on differences determined between classes of cells using any method of analyzing for the presence of genes, nucleic acids, or fragments thereof, including nucleic acid microchip arrays and differential display of expressed genes or nucleic acid fragments.

As used herein, "sample" relates to a particular experimental state for which all the variables being studied in a project are held fixed. By way of nonlimiting example, if a variable is a class of cell, the "sample" refers to a particular cell type; if a variable is the subsequence pairs employed in the project, a "sample" refers to a particular subsequence pair; or if a variable is a set of putative pharmaceutical agents, a "sample" refers to a particular agent from the set. As used herein, "representation" relates to any graphical, visual, or equivalent non-verbal display that provides an image of the results obtained according to the methods of the present invention. More specifically, a "representation" of the invention is obtained by transforming the quantitative results gathered by experiments underlying the invention. Examples of such data include, by way of non-limiting example, differential gene expression across classes of cell, and/or across a set of putative therapeutic agents, and/or equivalent types of experimental parameter.

In important embodiments, a representation of the invention is generated by algorithms executed in a computer and is suitable for display on a display means, such as a display screen or monitor, employed in the operation of the computer. The representation is also suitable for storing in a storage module or data archive of such a computer. It is still further suitable for printing from the computer onto a medium such as paper or equivalent physical medium, and for recording it onto a portable storage medium, including, for example, magnetic media, CD ROMs and equivalent storage media. As used

herein, "display means" includes any of the objects and media identified above in this paragraph, as well as equivalent apparatuses and objects suitable for displaying the results of computational processes for visual inspection.

As used herein, "extent of relatedness" is a characterization according to methods of the present invention of a degree of similarity or a degree of non-similarity between any two members of the same type of element; in particularly important embodiments, the type of element may be classes of cells.

As used herein, a "putative pharmaceutical agent" relates to a chemical compound or a composition comprising at least one chemical compound which is a candidate for being a therapeutic agent. Any such therapeutic agent may be used in treating a mammal suffering from a disease or a pathology. In treating the mammal with the therapeutic agent it is intended to attenuate the symptoms and/or the underlying causes of the disease or the pathology, to ameliorate the symptoms and/or the underlying causes, and/or to contribute to a cure of the disease or the pathology. Non-limiting examples of a putative pharmaceutical agent include an agent drawn from a chemical compound library; an isolate from a natural source; a compound synthesized specifically as a putative agent; or a substance derived or obtained using the practices of genetic engineering and recombinant nucleic acid technology such as a recombinant protein, a fragment of a recombinant protein, a recombinant polypeptide, a fragment of a recombinant polypeptide, a recombinant peptide, or a nucleic acid including, for example an oligonucleotide intended as an antisense agent, and a recombinant gene intended for administration as a gene therapeutic agent.

As used herein, a "fragment" of a nucleic acid relates to a contiguous portion originating from the genomic or cDNA-derived nucleic acid from a class of cells. The contiguous portion includes at or near each end a target subsequence defined according to the operational procedures disclosed herein, and includes all nucleotides in the sequence of the fragment bounded by the two target subsequences. The nucleotides between the two target subsequences, together with the subsequences themselves, define a "length" of the fragment, as used herein. The target subsequences are identified, for example, by contacting the nucleic acid from the cells with a specific pair of restriction endonucleases, or with a specific pair of oligonucleotide primers, and in equivalent ways.

The information used in the present invention is obtained from experiments providing the results of differential gene expression wherein the difference relates to an experimental state and a reference state. Commonly a reference state refers to a normal, or an unperturbed, or a non-pathological class of cells. An experimental state may relate to a certain set of conditions applied to one class of cells, and the corresponding reference state then relates to the same set of conditions applied to a second class of cells.

An experimental state may also relate to a class of cells in the presence of one or more putative therapeutic agents, in which case the reference state relates to the same class of cells in the absence of any putative therapeutic agent. An experimental state may furthermore be obtained from a class of cells that is of interest in a particular set of circumstances. This includes cells of a given cell type, cells from a given tissue, and cells from a given organ, and further includes cells that may be noncancerous or cancerous. Types of cell encompassed within the present invention include, by way of non-limiting example, endothelial cells, mesothelial cells, and epithelial cells. Tissues and organs included within the present invention may be, by way of non-limiting example, lung, heart, skeletal muscle, smooth muscle, brain, central nervous system, peripheral nervous system, stomach, liver, kidney, reproductive tissues and organs, skin, and bone. Cancerous cells include, by way of non-limiting example, cells from prostate cancer, breast cancer, colon cancer, lung cancer, lymphatic or hematopoietic cancers, and also include cells obtained from tissue biopsies or from cell lines in the National Cancer Institute human tumor cell line panel. The cells subjected to analysis in the present invention may also originate from plants, yeast, fungi, and other taxonomic groupings.

The methods of evaluating the extent of relatedness between classes of cells, for example, between a first class of cells and a second class of cells, are founded on evaluating the extent of relatedness of the expression of particular genes between the cells of the two classes. In a preferred embodiment of the invention, similarities and differences in the susceptibility of the nucleic acid present in the cells to digestion by specific pairs of restriction endonucleases are determined, according to the methods of the present invention, by procedures that are disclosed in detail in co-owned U. S. Patent No. 5,871,697 to Rothberg *et al.*, and in Shimkets *et al.* 1999 (Nature Biotechnology 17:798-803), both of which are incorporated herein by reference in their entirety.

Briefly, for any experimental state of a class of cells, the nucleic acid content of the cells, preferably in the form of a preparation of cDNA from the cells, is subjected to restriction endonuclease ("RE") digestion by specific pairs of endonucleases. Each member of the RE pair is chosen to optimize the likelihood that a restriction fragment resulting from the nuclease digestion will be a unique fragment. In an important implementation of this method, the restriction nuclease digestion is carried out on cDNA prepared from the cells of the class in the given experimental state. This implementation leads to emphasis on genes that are expressed in the experimental state, many of which may be characteristic of the given experimental state and be more poorly expressed, or not expressed at all significantly, in a different experimental state. A large number of specific pairs of nucleases may be employed. Alternatively, expression of a gene may be repressed in a characteristic way in a given experimental state and be expressed at a higher level, such as at a constitutive level, in a different experimental state. By

way of non-limiting example, several pairs of restriction nucleases that may be employed in implementing the present invention are disclosed in U. S. Patent No. 5,871,697.

In an alternative embodiment, the extent of relatedness may be obtained by amplification fragment length polymorphism analysis ("AFLP"). Briefly, amplification of the nucleic acid content of the class of cells being examined is subjected to a primer-dependent amplification procedure in which any of a set of primer pairs is used to initiate amplification. Amplification procedures are described in considerable detail in, for example, Innis *et al.*, PCR PROTOCOLS, A GUIDE TO METHODS AND APPLICATIONS, Academic Press, New York (1989), and Innis *et al.*, PCR STRATEGIES, Academic Press, New York (1995). The primers of each primer pair are different from each other, and reflect different subsequences that are the object of the amplification process. Amplification may proceed by any procedure, including polymerase chain reaction, known in the field of molecular biology. In AFLP, the length of an amplicon found in a given experimental state differs from the length found in a different experimental state. This may arise, for example, if the given experimental state arises from a mutation that occurs in a subsequence recognized by a primer used in the amplification reaction. It may also arise from a deletion from, or an insertion into, the nucleic acid of the cells in that state.

The experimental and computational procedures that may be employed to generate the representations of the present invention are described generally below.

Measurements

At the outset, the gene expression levels are determined experimentally. This can be done, in a preferred embodiment, by following the general protocols of differential expression using restriction endonucleases (U.S. Patent No. 5,871,697). For each pair of restriction enzymes and each biological sample, a pool of fluorescently-labeled DNA fragments is generated. Electrophoresis is then performed to separate these fragments based on size, and an intensity, designated as $I_{srt}(x)$, where s labels the sample, *i.e.*, the cell class; r labels the restriction enzyme pair, *i.e.*, the gene fragment; t labels the trial, and x is the length of the fragment as determined by electrophoresis, is detected. The length x may be either a continuous index or a convenient discretization. As an example, the resolution of the electropherogram may be set to a discretization of 0.1 nucleotide ("nt"). Commonly three independent trials are performed. A mean signal $I_{sr}(x)$ is then obtained by averaging over the n_t trials,

$$I_{sr}(x) = (1/n_t)\ \Sigma_t\ I_{srt}(x)\ \Sigma_1 \qquad (1)$$

Next, lengths x for each restriction enzyme pair r where some of the samples have a significant difference in measured intensity are identified. Such a difference is determined with respect to cell types, or with respect to the presence vs. the absence of a putative pharmaceutical agent. Labeling the d^{th} such difference d, the values $I_{sd} = I_{sr}(x)$ are then collected. Any of several methods for identifying significant differences may be employed, some of which are outlined herein. For example, an important method involves the following computational steps:

1. The mean $I_r(x) = \Sigma_s I_{sr}(x)$ is evaluated.
2. All positions, *i.e.* lengths, where, for at least one sample, $I_{sr}(x) - I_r(x)$ is larger than some threshold value, are marked.
3. The largest value of $I_{sr}(x) - I_r(x)$, determined as a difference between a sample state and the mean for restriction enzyme pair r, is found and the length x, indexing the difference, is marked.
4. Step 3 is repeated for succeedingly smaller values of the intensity difference. If the length x that marks the current largest difference is within a distance w from the length of a previously identified difference, the current difference is skipped and the next smaller difference is considered.
5. Step 4 is repeated until there are no more differences to consider.

Another method involves finding differences that meet a statistical criterion. A particular example of such a method involves the computational steps of:

1. defining a set of sample classes and assigning each sample to a particular class c;
2. for each restriction enzyme pair r and length x, evaluating the F-statistic for the set of measurements $I_{sr}(x)$ and the classes c to which samples are assigned, thereby providing the probability $p_r(x)$ that any differences between sample classes may be explained by random variation (See, for example, P. Hinton, Statistics Explained, Routledge 1995);
3. ordering the probabilities $p_r(x)$ from smallest (most significant) to largest (least significant);
4. optionally truncating the list at some threshold value of $p_r(x)$ above which differences are no longer considered significant (accepted values are $p_r(x) = 0.01$ to 0.05);
5. finding the smallest value of $p_r(x)$ and marking the length x as a difference for restriction enzyme pair r;

6. repeating step 4 and determining whether the length x that marks the current difference is in a region that is within a distance w of a previous difference, in which case the current difference is skipped and the next smaller distance is considered; and

7. continuing until there are no more differences to consider.

These exemplary computational procedures provide a set of measures of intensity I_{sd} for the class of cells in sample s at difference d.

Distances

For hierarchical clustering, a distance $D_{ss'}$ may be defined as the distance in vector space between pairs of samples s and s'. A variety of methods for calculating $D_{ss'}$ are available. Some examples, which are intended as being nonlimiting, are provided below.

$D_{ss'}$ as a scaled correlation function:

1. One calculates $\mu_d = (1/n_s)\, \Sigma_s I_{sd}$ and $\sigma_d = [(1/n_s)\Sigma_s (I_{sd} - \mu_d)^2]^{0.5}$. If data is missing, for example no measurement of I_{sd} exists for some sample s, that sample is excluded from the sum and n_s is reduced by 1.

2. One calculates $J_{sd} = (I_{sd} - \mu_d) / \sigma_d$. If data is missing for I_{sd}, then J_{sd} is defined as $J_{sd} = 0$.

3. One calculates $\mu_s = (1/n_d)\, \Sigma_d J_{sd}$ and $\sigma_s = [(1/n_d)\Sigma_d (J_{sd} - \mu_s)^2]^{0.5}$.

4. One calculates $K_{sd} = (J_{sd} - \mu_s) / \sigma_s$.

5. One calculates the covariance matrix $S_{ss'} = (1/n_d)\, \Sigma_d K_{sd} K_{s'd}$

6. One calculates the correlation matrix $C_{ss'} = S_{ss'} / [S_{ss} S_{s's'}]^{0.5}$.

7. One calculates $D_{ss'} = [2 - 2\, C_{ss'}]^{0.5}$.

$D_{ss'}$ as a Euclidean distance: $D_{ss'} = [\, \Sigma_d (I_{sd} - I_{s'd})^2\,]^{0.5}$.

$D_{ss'}$ as a Pearson distance: $D_{ss'} = [\, \Sigma_d (I_{sd} - I_{s'd})^2 / \sigma_d^2\,]^{0.5}$ where σ_d is defined in step 1 of scaled correlation function above.

$D_{ss'}$ as a pairwise Pearson distance:

1. One calculates the covariance matrix $S_{ss'} = (1/n_d)[\, \Sigma_d I_{sd} I_{s'd} - (\Sigma_d I_{sd})(\Sigma_d I_{s'd}) / n_d \,]$.

2. One calculates the correlation matrix $C_{ss'} = S_{ss'} / [\, S_{ss} S_{s's'} \,]^{0.5}$.

3. One calculates $D_{ss'} = [\, 2 - 2\, C_{ss'} \,]^{0.5}$.

$D_{ss'}$ as a Mahalanobis distance:

1. One calculates the covariance matrix $S_{dd'} = (\Sigma_s I_{sd} I_{sd'}) - (\Sigma_s I_{sd})(\Sigma_s I_{sd'})/n_s$

2. One calculates the correlation matrix $C_{dd'} = S_{dd'} / [\, S_{dd} S_{d'd'} \,]^{0.5}$ and its matrix inverse $C^{-1}_{dd'}$.

3. One calculates $D_{ss'} = [\, \Sigma_{dd'} (I_{sd} - I_{s'd})\, C^{-1}_{dd'}\, (I_{sd'} - I_{s'd'}) \,]^{0.5}$

It is contemplated that other distance methods known in the art may be used in the invention, such as Spearman correlation, and the like. Other methods known in the art can be found, for example and not be means of limitation, in K. V. Mardia, J. T. Kent, and J. M. Bibby, MULTIVARIATE ANALYSIS, Academic Press, New York, 1979.

Hierarchical Clustering

The distances can be used to perform hierarchical clustering of the samples. A general algorithm for clustering is described below.

1. Each sample s is assigned its own initial cluster c.

2. One calculates all the distances between pairs of clusters and finds the smallest distance. These two clusters are joined into a single cluster and the number of clusters is decreased by 1.

3. Step 2 is repeated until only a single cluster remains.

In order to implement this algorithm, a method to calculate the distance between pairs of clusters is also required. Some nonlimiting examples of such calculations, using well-known methods, are indicated below.

Nearest neighbor, single linkage: The distance between clusters c and c' is the smallest distance $D_{ss'}$, where s ranges over all samples in cluster c and s' ranges over all samples in cluster c'.

Unweighted pair group method using arithmetic averages (UPGMA), also known as average linkage: The distance between clusters c and c' is $(\Sigma_{ss'} D_{ss'}) / (n_c n_{c'})$ where s ranges over all samples in

cluster c, s' ranges over all samples in cluster c', n_c is the number of samples in cluster c, and $n_{c'}$ is the number of samples in cluster c'.

Furthest neighbor, complete linkage: The distance between clusters c and c' is the largest distance $D_{ss'}$ where s ranges over all samples in cluster c and s' ranges over all samples in cluster c'.

Other distance-based hierarchical clustering methods are well-known. See, for example, Wen-Hsiung Li, MOLECULAR EVOLUTION, Sinauer Assoc, 1997.

Software packages are available to perform the clustering and display the results. See, for example, Phylip, Joe Felsenstein, http://evolution.genetics.washington.edu for clustering, and Treeview, Rod Page, http://taxonomy.zoology.gla.ac.uk/rod/treeview.html for display. The source code for the unit within Phylip employed for the clustering, and the downloaded executable file of Treeview for Windows 95 and Windows NT, as well as a manual for Treeview, are available from the owner of the present application.

Two-Dimensional Clustering

It is also possible to cluster the distances, rather than clustering the samples. One simply exchanges the roles of the samples and differences in the equations above. Furthermore, it is possible to perform clustering of both samples and differences, and then to display the measurements I_{sd} in which both samples and differences are presented in cluster order.

Principal Component Analysis and Principal Factor Analysis

Principal component analysis is described in standard texts. See, for example, Mardia, Kent, and Bibby. To perform principal component analysis, one begins with a correlation matrix $C_{ss'}$ as defined above, in the section "Distances". (Alternatively, one could use the covariance matrix $S_{ss'}$). Eigenvalues and eigenvectors, defined such that $C_{ss'} g_{s'i} = a_i g_{si}$, where the i^{th} eigenvalue is a_i and its eigenvector is g_{si}, are calculated. The eigenvalues are ordered from largest to smallest: $a_1 \geq a_2 \geq \ldots \geq a_S$. To obtain a reduced dimensional depiction of the samples, a number of desired dimensions k is chosen. Then, in k-dimensional space, sample s is represented as the point $(g_{s1}, g_{s2}, \ldots, g_{sk})$. Samples that are close in the k-dimensional space have similar expression profiles and may be considered to be related.

As an alternative to using the correlation matrix $C_{ss'}$ as the starting point for principal component analysis, it is possible to calculate principal components using the inner product matrix from multidimensional scaling defined as

$$B = H\,C\,H \qquad (2)$$

where C is the correlation matrix, H is the centering matrix with diagonal elements given by 1 - (1/n) and off-diagonal elements - (1/n), where n is the number of items being correlated. (See, for example, Mardia, Kent, and Bibby, Multivariate Analysis, and Arkin, Shen, and Ross, Science 277: 1275 (1997)). The k^{th} principal component is then the k^{th} eigenvector of B normalized to unit length and ordered by decreasing eigenvalue λ_k, and the k^{th} principal factor is obtained by scaling the eigenvector by $\lambda_k^{1/2}$. The projection of sample s onto the k^{th} principal factor is the element of the factor for row s. The components or factors are ordered from 1 (corresponding to the most informative) to n (corresponding to the least informative). By using some, but not all, of the components or factors, the samples can be represented in a small-dimensional geometric space. Furthermore, the amount of information retained in the representation can be related to the eigenv.....es of the components that are used (See Mardia, Kent, and Bibby).

A centered inner product matrix B appropriate for principal component or prinicpal factor analysis can also be obtained from any distance matrix $D_{ss'}$ as

$$B = H\,A\,H \qquad (3)$$

where

$$A_{ss'} = -1/2\,(D_{ss'})^2. \qquad (4)$$

To perform principal factor analysis, factor i is defined as $h_{si} = a_i^{0.5}\,g_{si}$ where, as before, a_i is the eigenvalue of the i^{th} eigenvector g_{si}. An orthonormal rotation matrix G ($\Sigma_j G_{ij}G_{kj}$ is 1 if i = k and 0 otherwise, det(G) = +1) is introduced and the factors are rotated to obtain rotated coordinates for the samples. Thus, to obtain a k-dimensional representation of the locations of the samples, the following operations are performed:

1. One calculates the correlation matrix $C_{ss'}$ or the covariance matrix $S_{ss'}$ where s and s' label individual samples.

2. One calculates the eigenvalues a_i and eigenvectors g_{si} for the matrix, with $a_1 \geq a_2 \geq \ldots \geq a_n$.

3. Unrotated factor loadings $h_{si} = a_i^{0.5}\,g_{si}$ are defined.

4. The first k factor loadings and an orthonormal rotation matrix G are selected. The j^{th} coordinate of sample s in the rotated space is $\Sigma_i\ h_{si}\,G_{ji}$

The rotation matrix G may be optimized according to standard criteria. See, for example, Mardia, Kent, and Bibby, Ch. 9.6 on Varimax rotation, *supra*. The rotated axes represent factors that influence the observed measurements for the samples

In implementing the methods of the present invention, these operations may be sequentially combined in any of several ways according to the intended display, *i.e.*, the nature of the relatedness that is intended to be shown.

Also, the information from the principal factors can be used to help filter the experimental noise from the correlation functions. For example, it is possible to select a cut-off principal factor $j < n$, then compute distances and correlations between samples based on their representation in the j-dimensional principal factor space.

As a nonlimiting example of the computational procedures that may be employed in the present invention, a schematic overview of procedures that may be adopted is presented in Figure 1. The experimental results represent the sample-dependent and selection-dependent intensities obtained in an experiment, arrayed in a measurement matrix. In the implementation shown in Figure 1, the difference bands having various, defined, nucleotide lengths are arrayed as the columns of the matrix; they are obtained in various experiments that are selected using different members of the sets of subsequence pairs. The samples represent the classes of cells, or cells treated with a set of putative pharmaceutical agents, or analogous sample sets, and are arrayed as the rows.

The values arrayed in the measurement matrix may then be subjected to correlation analysis to provide either direct sample correlations or correlations of differences. The measurement matrix can also be subjected to a calculation providing a vectoral distance between samples; such a sample distance may also be obtained from the sample correlation result. The distance vector can further be subjected to a linkage analysis to provide hierarchical clustering of the samples. Additionally, the correlated samples may be subjected to principal component analysis providing the principal factors contributing to a state or to a difference

A nonlimiting example of the way in which a principal component analysis may be carried out, using methods described herein, is presented in Figure 2. The correlation matrix or the centered inner product matrix described above is subjected to appropriate operations to provide the principal components and the principal factors, based on their eigenvalues and eigenvectors. Advantageously a reduction in the number of dimensions employed in the number of eigenstates may provide a filtering effect, reducing the noise in the vector distances calculated.

The representations provided in the present invention find use in various applications of genomics in the biological and medical fields. Extents of relatedness and correlations provide rapid overviews of enzymatic reactions, metabolic pathways, and physiological effects that become distinguished when comparing states. When a pathological state is compared with a normal state, for

example in a mammal, and especially in a human, the display of distinguished pathways is instructive in the development of therapeutic approaches and/or therapeutic agents for the treatment of the pathological state. When a putative pharmaceutical agent is compared to a state that omits the agent, or when one such agent is compared with another, important information is provided relating to the metabolic reactions induced by or undergone by the agent or agents, leading to optimal choice of such agents. This information may also provide leads to the development of novel pharmaceutical agents. If the genome being studied is a plant genome, such as the genome of an important crop plant, analogous principles apply.

Nucleic acid assays

The present invention provides a method for generating a representation of the extent of relatedness between at least two classes of cells. In this method, the cells in each class are chosen from among cells of a given cell type, cells from a given tissue, and cells from a given organ. Generation of nucleic acids from the cell samples of choice may be as described in the GeneCalling™ methodology. See U.S. Patent No. 5,871,697. The method includes the steps of: (a) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence; (b) isolating the nucleic acid of each class of cells and assaying for the presence of a nucleic acid fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and quantitating the extent to which each fragment is present; and (c) determining the extent of relatedness reflecting similarities or differences in the presence and quantitation of the fragments among the classes using software algorithm programs known in the art.

One important embodiment of this method, *i.e.*, determining the presence of the fragments and quantitating the amounts present, as described in step (b) above, is carried out by a process that includes the steps as follow. First, samples of the nucleic acid from the cells of each class are digested with a plurality of specific pairs of restriction endonucleases ("REs"). Each sample is treated by one RE pair, where one RE of the pair targets the first subsequence described in step (a) above, and the second RE of the pair targets the second subsequence, with each digestion providing specific restriction fragments.

Second, double stranded adapter DNA molecules are hybridized to the fragments. Each adapter DNA molecule comprises: (*i*) a shorter strand, preferably having no 5' terminal phosphate, consisting of a first and second portion, the first portion being a region at the 5' end that is complementary to the overhang produced by one of the REs of the given pair and a second portion hybridizable to the opposite longer strand of the adaptor, and (*ii*) a longer strand, preferably having no 5' terminal phosphate,

consisting of a first portion at its 3' end complementary to the above-mentioned second portion of the shorter strand, and an optional second portion at its 5' end comprising a unique region not hybridizable to any sequence present in the original sample population. See U.S. Patent No. 5,871,697. The longer strand is optionally labeled with fluorochrome 208, although any DNA labeling system that preferably allows multiple labels to be simultaneously distinguished is usable in this invention. See, *e.g.*, Ausubel, *et al.* CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, New York, NY, 1993.

Third, output signals from each ligated fragment are detected for each sample population so treated. Each ligated fragment generates output signals that characterize (a) the presence of the given subsequences corresponding to the RE pair used in a particular run, (b) the length between the two subsequences corresponding to the two REs employed in a given run, and (c) the quantitation of the relative amounts present of each fragment so generated in a given run.

Optionally, a nucleotide sequence database may be searched for sequences that are predicted to produce, or alternatively, not produce, the one or more output signals generated by the nucleic acid from the cells of each class, given the parameters described above. The analysis methods comprise, first, selecting a database of DNA sequences representative of the DNA sample to be analyzed, second, using this database and a description of the experiment to derive the pattern of simulated signals that would be generated, contained in a database of simulated signals, that will be produced by DNA fragments generated in the experiment, and third, for any particular detected signal, using the pattern or database of simulated signals to predict the sequences in the original sample likely to cause this signal. Further analysis methods present an easy to use user interface and permit determination of the sequences actually causing a signal in cases where the signal may arise from multiple sequences, and perform statistical correlations to quickly determine signals of interest in multiple samples. A sequence from a searched database is predicted to produce the one or more output signals when that sequence has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (b) the same target nucleotide sub-sequences that are represented by said one or more output signals, or target nucleotide subsequences that are members of the same sets of target nucleotide sub-sequences represented by the one or more output signals.

A first analysis method is selecting a database of DNA sequences representative of the sample to be analyzed. In the preferred use of this invention, the DNA sequences to be analyzed will be derived from a tissue sample, typically a human sample examined for diagnostic or research purposes. In this use, database selection begins with one or more publicly available databases which comprehensively record all observed DNA sequences. Such databases are GenBank from the National Center for

Biotechnology Information (Bethesda, Md.), the EMBL Data Library at the European Bioinformatics Institute (Hinxton Hall, UK) and databases from the National Center for Genome Research (Santa Fe, N.Mex.). However, as any sample of a plurality of DNA sequences of any provenance can be analyzed by the methods of this invention, any database containing entries for the sequences likely to be present in such a sample to be analyzed is usable in the further steps of the computer methods.

A second analysis method uses the previously selected database of sequences likely to be present in a sample and a description of an intended experiment to derive a pattern of the signals which will be produced by DNA fragments generated in the experiment. This pattern can be stored in a computer implementation in any convenient manner. In the following, without limitation, it is described as being stored as a table of information. This table may be stored as individual records or by using a database system, such as any conventionally available relational database. Alternatively, the pattern may simply be stored as the image of the in-memory structures which represent the pattern.

A second important embodiment of this method, *i.e.*, determining the presence of the fragments and their quantitation, as described in step (b) above, is carried out by a process that includes the steps as follow. First, for each pair of nucleotide subsequences selected, a pair of oligonucleotide primers are provides, the pair consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence. Second, the nucleotide sequence between the first subsequence and the second subsequence are amplified using the oligonucleotide primers to prime the amplification, thereby providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to the two primers employed in each pair and a quantitation of the extent to which each amplicon is present. Third, output signals are generated as above for each amplicon, each output signal characterizing (a) the subsequences of the pairs of primers, (b) the length, and (c) the quantitation. Optionally, a nucleotide sequence database may be searched for sequences that are predicted to produce, or alternatively, not produce, the one or more output signals generated by the nucleic acid from the cells of each class, given the parameters described above. Analysis methods are as described above.

This invention can be applied, for example and not by way of limitation, to *in vitro* cell populations or cell lines, to *in vivo* animal models of disease or other processes, to human samples, to purified cell populations perhaps drawn from actual wild-type occurrences, and to tissue samples containing mixed cell populations. The cell or tissue sources can advantageously be a plant, a single celled animal, a multicellular animal, a bacterium, a virus, a fungus, or a yeast, etc. The animal can

advantageously be laboratory animals used in research, such as mice engineered or bred to have certain genomes or disease conditions or tendencies.

Cells used in the invention may be obtained from a mammal, preferably a human, having or suspected of having a diseased condition. In one embodiment, the diseased condition is a malignancy. The *in vitro* cell populations or cell lines can be exposed to various exogenous factors to determine the effect of such factors on gene expression. In a preferred embodiment, the exogenous factor is a putative pharmaceutical agent. Cells so contacted with a putative pharmaceutical agent are treated with an amount of the agent sufficient to effect a change in the state of those cells or with an amount of the agent less than or equal to a predetermined upper limit of dosing concentration, prior to their being assayed. Measures of relatedness and extent of correlation may be made between cells so contacted with putative pharmaceutical agent and, for example, cells not so contacted.

Extent of relatedness methodology

The present invention provides a representation of the extent of relatedness between a first class of cells and a second class of cells. The cells in each class are chosen from among cells of a given cell type, cells from a given tissue, and cells from a given organ, as described above. The extent of relatedness reflects similarities or differences in the presence of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence, in a nucleotide length separating the first and second subsequences of the pair and in a quantitation of the extent to which each pair having the determined length is in the classes of cells. Input information of the fragments to be analyzed are obtained by methods of nucleic acid analysis and quantitation as described in the NUCLEIC ACID ASSAYS section above.

The measure of relatedness is provided by calculating a distance that reflects the amplitude of a difference vector. A difference vector is defined as a difference between a first vector and a second vector. Herein, the first vector reflects information derived from the quantitation for each subsequence pair obtained for the first class of cells, and correspondingly, the second vector reflects the analogous information derived from the second class. The different elements of each vector relate to data obtained using different subsequence pairs.

In an embodiment of the representation, the extent of relatedness is related to a distance. This distance reflects the amplitude of a difference vector that is a difference between a first vector which reflects information derived from the quantitation for each subsequence pair obtained for the first class and a second vector which reflects the corresponding information obtained for the second class. The different elements of each vector relate to data obtained using different subsequence pairs.

In an additional significant embodiment, the representation includes a tree structure reflecting the extent of relatedness is provided by generating a tree structure reflecting the relatedness between any two classes. The branches of the tree structure reflect the difference vectors and are ramified from nodes.

In important embodiments of the representation of the extent of relatedness, the representation is obtained employing the methods of the invention, including the methods that have been summarized in the paragraphs immediately above.

In additional significant embodiments of the representation of the extent of relatedness, the cells in at least one class are obtained as described in the NUCLEIC ACID ANALYSIS section above.

Correlation analysis methodology

The invention also provides a method for generating a representation of the correlation between a first class of cells and a second class of cells. The correlation reflects a change in the nature and amount of nucleic acids present in the classes. In this method, the cells in each class are chosen from among cells of a given cell type, cells from a given tissue, and cells from a given organ. The method of nucleic acid analysis and quantitation are as describe in the NUCLEIC ACID ASSAYS section above.

Upon generation of a signal output, the correlation between the cells of the first class and cells of the second class are correlated, and a representation of the correlation is prepared. The quantitation of the fragments in the invention corresponding to the RE pair used in a given run and the length of each fragment so generated; thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

In a significant embodiment of the method for generating a representation of the correlation, the correlation is related to a set of orthonormal eigenvectors, as described in the DISTANCES section above. The elements of the basis set upon which the eigenvectors are constructed reflect particular biochemical or physiological pathways correlated between the cells of the two classes. Each of these eigenvectors is associated with an eigenvalue that is an integer greater than zero. After defining an upper limit of the eigenvalues to be used, the coefficients of the basis set elements in each eigenvector whose eigenvalue is less than or equal to this upper limit reflects the contribution of the corresponding pathway to the biochemical or physiological differences correlated between the cells of the first class and the cells of the second class.

In another significant embodiment of the method for generating a representation of the correlation, the representation is a cluster diagram or a dendrogram, and includes a tree structure reflecting the relatedness of the pathways involved in the biochemical or physiological response to a difference between cells of the two classes. In obtaining this representation, a correlation matrix is calculated that provides a distance determination in which the distance reflects the amplitude of a difference vector. This vector is a difference between two vectors each of which reflects information obtained for the response of one of the two classes to the difference between the classes, and wherein the branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

In additional significant embodiments of the representation of the extent of correlation, the cells in at least one class obtained as described in the NUCLEIC ACID ANALYSIS section above.

Display means

The present invention also provides a display means displaying a representation of the extent of relatedness between a first class of cells and a second class of cells. The cells in each class are chosen from among cells of a given cell type, cells from a given tissue, and cells from a given organ, as described above. The extent of relatedness reflects similarities or differences in the presence of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence, in a nucleotide length separating the first and second subsequences of the pair and in a quantitation of the extent to which each pair having the determined length is in the classes of cells.

In a significant embodiment of the display means, the extent of relatedness is related to a distance. This distance reflects the amplitude of a difference vector that is a difference between a first vector which reflects information derived from the quantitation for each subsequence pair obtained for the first class and a second vector which reflects the corresponding information obtained for the second class. The different elements of each vector relate to data obtained using different subsequence pairs.

In an additional significant embodiment of the display means, the representation includes a tree structure reflecting the relatedness between any two classes, in which the branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

In important embodiments of the display means displaying a representation of the extent of relatedness, the representation is obtained employing the methods of the invention, including the methods that have been summarized in the paragraphs immediately above.

In additional significant embodiments of the display means displaying a representation of the extent of relatedness, the cells in at least one class obtained as described in the NUCLEIC ACID ANALYSIS section above.

The present invention additionally provides a display means displaying a representation of the correlation between a first class of cells and a second class of cells. The cells in each class are chosen from among cells of a given cell type, cells from a given tissue, and cells from a given organ, as described above. The correlation reflects differences between the first class and the second class in the presence of a pair of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence and the nucleotide length separating the first and second subsequences of the pair, and in a quantitation of the extent to which each pair having the determined length is present in the cells.

In an advantageous embodiment of this display means, the correlation is related to a set of orthonormal eigenvectors. The elements of the basis set upon which the eigenvectors are constructed reflect particular biochemical or physiological pathways correlated between the cells of the two classes Each of these eigenvectors is associated with an eigenvalue that is an integer greater than zero. After defining an upper limit of the eigenvalues to be used, the coefficients of the basis set elements in each eigenvector whose eigenvalue is less than or equal to this upper limit reflect the contribution of the corresponding pathway to the biochemical or physiological differences correlated between the cells of the first class and the cells of the second class.

In an additional advantageous embodiment of the display means displaying a representation of the correlation, the representation is a cluster diagram or a dendrogram, and includes a tree structure reflecting the relatedness of the pathways involved in the biochemical or physiological response to a difference between cells of the two classes. In obtaining this representation, a correlation matrix is calculated that provides a distance determination in which the distance reflects the amplitude of a difference vector. This vector is a difference between two vectors each of which reflects information obtained for the response of one of the two classes to the difference between the classes. The branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

In important embodiments of the display means displaying a representation of the correlation, the representation is obtained employing the methods of the invention, including the methods that have been summarized in the paragraphs immediately above.

In important embodiments of the representation of the correlation, the representation is obtained employing the methods of the invention, including the methods that have been summarized in the paragraphs immediately above.

In additional significant embodiments of the display means displaying a representation of the correlation, the cells in at least one class obtained as described in the NUCLEIC ACID ANALYSIS section above.

Other Aspects

In addition to providing representations of cells, the techniques described here are also useful for providing representations of nucleic acid fragments or genes. The starting point for the analysis is the matrix I_{sd} described previously, where s labels the sample (or group of samples or distinct types of cells) and d labels a particular measurement of the expression level of a particular gene in that class. Rather than generating representations based on the rows of I, each representing a different sample or group of samples, it is possible to generate representations based on the columns of I, each representing a different nucleic acid. Hierarchical and geometrical representations of nucleic acids, based on their relative abundance across a series of cells, can be used to infer genes that are co-expressed and are likely to have related biological function.

Other Embodiments

The data matrix of intensities I can be described more generally as a representation in which each row corresponds to a particular biological sample or group of samples, and each column corresponds to a particular nucleic acid molecule or class of molecules whose quantities are measured in each of the biological states.

In addition to the differential-display methods described to provide measurements of nucleic acid quantities, other methods for obtaining measurements of the nucleic acids present in a cell are available. These include restriction fragment length polymorphism, amplification fragment length polymorphism, EST sequencing, serial analysis of gene expression, hybridization to oligonucleotide probes, and other methods known in the art. Other methods, such as quantification by TaqMan or Northern blots, are also used. All of these methods generate data sets that can be analyzed according to the methods described here. The measurements I_{sd} for each biological state and nucleic acid can correspond to absolute concentrations, concentrations relative to a standard (either ratio or numeric difference), or other convenient measures.

The methods of the invention includes analysis of populations ranging from 5, 10, 25, 50, 100, 1000, 10,000 or 100,000 or more members.

EXAMPLE

Male Sprague-Dawley rats (Harlan Sprague Dawley, Inc., Indianapolis, Indiana) of 10-14 weeks of age were gavage-fed and dosed once a day for three days with the following drugs, dissolved in sterile water, at the following levels:

phenobarbitol	3.81 mg/kg/day
gabapentin	34.29 mg/kg/day
vigabatrin	150 mg/kg/day
paraldehyde	77.08 mg/kg/day.

These dosages correspond to the ED100 (the upper limit of the effective dose for humans) adjusted for the difference in metabolic rate between rats and humans. Three rats were used for each drug treatment, and an additional three rats to match each drug were treated with sterile water to serve as a control.

Rats were sacrificed 24 hours after the final dose and their brains were harvested. Collection of mRNA, synthesis of cDNA, and differential display protocols were carried out according to methods described in U. S. Patent No. 5,871,697 and Shimkets *et al.* 1999 (Nature Biotechnology 17:798-803).

The following steps were followed to analyze the differential display pattern:

1. The intensities $I_{ai}(x)$ for each of the three animals treated with the same drug were combined into a single average $I_{ar}(x)$, where the subscript a labels the drug. The standard deviation $s_{ar}(x)$ was also computed for the measurements from the individual animals treated with the drug.

2. The averages $I_{ar}(x)$ and standard deviations $s_{ar}(x)$ for each drug were compared with the average $I_{cr}(x)$ and standard deviation $s_{cr}(x)$ for the sterile water control treatment. A difference at length x was marked if

$$\mathrm{ABS}(\ln [I_{ar}(x)/I_{cr}(x)]) \geq \ln(1.5) \qquad (5)$$

and if the significance was smaller than 0.15 for a two-tailed t-test with

$$t = [I_{ar}(x) - I_{cr}(x)] / [\{ s_{ar}(x)^2 + s_{cr}(x)^2 \} / 2]^{1/2} \qquad (6)$$

and infinite degrees of freedom. The difference intensities marked according to this procedure may then be inspected by eye and visually significant differences may be retained.

3. For each of the differences d, defined by a restriction enzyme pair r and a position x, the intensity $I_{ar}(x) = I_{ad}$ was determined for each of the drug treatments, whether or not that particular treatment has a difference compared to the control.

In this example, the final data matrix I_{ad} has 8 rows: 1 row for each of the 4 drugs, and 1 row for each of 4 replicates of the water control data. The matrix has as many columns as the number of differences detected in the differential display pattern.

The Pearson correlation coefficient C_{ab} between the 8 classes of samples (4 drugs, 4 water controls) was determined using methods provided in the Detailed Description of the Invention. If a data element for a particular difference was missing for a particular treatment, that difference did not contribute to the correlation coefficient. The correlations are shown in the Table 1 below, with the standard deviation within a drug shown as the diagonal elements.

Table 1.

	vigabatrin	phenobarbitol	gabapentin	paraldehyde	water_1	water_2	water_3	water_4
vigabatrin	261.3601	0.9982	0.9913	0.9728	0.6549	0.6673	0.3674	0.3786
phenobarbitol	0.9982	263.4469	0.9973	0.9325	0.5675	0.4724	0.6569	0.6601
gabapentin	0.9913	0.9973	556.8114	0.9922	0.6177	0.6057	0.3400	0.2704
paraldehyde	0.9728	0.9325	0.9922	423.1916	0.6485	0.5307	0.5952	0.5702
water_1	0.6548	0.5578	0.6177	0.6485	59.4573	0.9718	0.9896	0.9735
water_2	0.6673	0.4724	0.6057	0.5307	0.9718	62.3568	0.9960	0.9886
water_3	0.3674	0.6569	0.3400	0.5952	0.9896	0.9960	107.2630	0.9978
water_4	0.3786	0.6601	0.2704	0.5702	0.9735	0.9886	0.9978	123.0743

Next the pairwise Pearson distance was calculated as described previously. The distance matrix is shown in Table 2 below.

Table 2.

	vigabatrin	phenobarbitol	gabapentin	paraldehyde	water_1	water_2	water_3	water_4
vigabatrin	0.0000	0.0597	0.1319	0.2333	0.8309	0.8157	1.1248	1.1148
phenobarbitol	0.0597	0.0000	0.0741	0.3675	0.9297	1.0272	0.8284	0.8245
gabapentin	0.1319	0.0741	0.0000	0.1245	0.8744	0.8883	1.1489	1.2080
paraldehyde	0.2333	0.3675	0.1245	0.0000	0.8384	0.9688	0.8997	0.9271
water_1	0.8309	0.9297	0.8744	0.8384	0.0000	0.2376	0.1446	0.2302
water_2	0.8157	1.0272	0.8830	0.9688	0.2376	0.0000	0.0895	0.1509
water_3	1.1248	0.8284	1.1489	0.8997	0.1446	0.0895	0.0000	0.0663
water_4	1.1148	0.8245	1.2080	0.9271	0.2302	0.1509	0.0663	0.0000

The distances were then used as input to a nearest-neighbor clustering algorithm. The resulting clusters, using sterile H_2O as an outgroup, was shown in Fig. 3. The horizontal distances in Fig. 1A were proportional to the pairwise Pearson distance between clusters.

The correlation matrix C_{ab} also served as the starting point for principal factor analysis. First, principal components were calculated using the inner product matrix from multidimensional scaling

$$B = H\,C\,H \qquad (2)$$

where C is the correlation matrix and H is the centering matrix. The k^{th} principal component is then the k^{th} eigenvector of B normalized to unit length and ordered by decreasing eigenvalue λ_k, and the k^{th} principal factor was obtained by scaling the eigenvector by $\lambda_k^{1/2}$. Projections of the treatments and controls onto principal factors are shown in Table 3 below.

Table 3.

factor	1	2	3	4	5	6	7	8
eigenvalue	1.841	0.347	0.093	0.036	0.007	0.000	-0.036	-0.389
vigabatrin	-0.513	-0.163	-0.099	0.083	0.004	0.000	0.000	0.000
phenobarbitol	-0.397	0.317	-0.132	-0.038	-0.019	0.000	0.000	0.000
gabapentin	-0.580	-0.157	0.028	-0.094	-0.004	0.000	0.000	0.000
paraldehyde	-0.404	0.127	0.193	0.057	0.034	0.000	0.000	0.000
water_1	0.368	-0.169	0.109	0.017	-0.060	0.000	0.000	0.000
water_2	0.422	-0.300	-0.091	-0.017	0.033	0.000	0.000	0.000
water_3	0.542	0.156	0.045	-0.090	0.012	0.000	0.000	0.000
water_4	0.561	0.190	-0.059	0.082	0.001	0.000	0.000	0.000

The components are ordered from 1 (most informative) to 8 (least informative). The negative eigenvalues arise from the method used to account for missing data. If missing data had been handled in an alternate manner, for example if a missing element had been set to the average value or if the analysis were restricted to differences for which no data was missing, the eigenvalues would all be non-negative.

In Fig. 4, the treatments are displayed by projection onto principal factors. Factor 1 discriminates between drugs, where it has a negative value, and controls, where it has a positive value. Factor 2 discriminates between the drug treatments.

EQUIVALENTS

From the foregoing detailed description of the specific embodiments of the invention, it should be apparent that unique methods for representing the extent of relatedness between cells, cell lines, tissues, organs, or expressed sequences based on a genomic analysis of gene expression have been described. Although particular embodiments have been disclosed herein in detail, this has been done by way of example for purposes of illustration only, and is not intended to be limiting with respect to the scope of the appended claims which follow. In particular, it is contemplated by the inventor that various substitutions, alterations, and modifications may be made to the invention without departing from the spirit and scope of the invention as defined by the claims. For instance, the choice of source material, subsequences used, or software algorithm used is believed to be a matter of routine for a person of ordinary skill in the art with knowledge of the embodiments described herein.

CLAIMS

I claim:

1. A method for generating a representation of the extent of relatedness between at least two classes of cells, wherein the cells in each class are chosen from the group consisting of cells of a given cell type, cells from a given tissue, and cells from a given organ, the method comprising the steps of

a) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

b) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present; and

c) determining the extent of relatedness reflecting similarities or differences in the presence and quantitation of the fragments among the classes.

2. The method described in claim 1 wherein the determining of the presence and quantitation of the fragments described in step b) is carried out by a process comprising the steps of:

i) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (a) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (b) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

ii) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (a) the subsequences of the pairs of restriction endonucleases (b) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (c) the quantitation of the fragment corresponding to the pair and the length; and

iii) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (b) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

3. The method described in claim 1 wherein the determining of the presence of the fragments and the quantitation of the fragments, described in step b) is carried out by a process comprising the steps of:

i) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

ii) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

iii) generating output signals for each amplicon, each output signal characterizing (a) the subsequences of the pairs of primers, (b) the length, and (c) the quantitation; and

iv) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (a) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals.

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

4. The method described in claim 1 wherein the extent of relatedness in step c) is provided by calculating a distance wherein the distance reflects the amplitude of a difference vector that is a difference between a first vector which reflects information derived from the quantitation for each subsequence pair obtained for the first class and a second vector which reflects information derived from the quantitation for each subsequence pair obtained for the second class, wherein different elements of each vector relate to data obtained using different pairs.

5. The method described in claim 1 wherein the extent of relatedness in step c) is provided by generating a tree structure reflecting the relatedness between any two classes, wherein the branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

6. The method described in claim 1 wherein the cells in at least one class are cancer cells.

7. The method described in claim 1 wherein the cells in at least one class have been contacted with a putative pharmaceutical agent.

8. A method for generating a representation of the correlation between a plurality of classes of cells wherein the cells in each class are chosen from the group consisting of cells of a given cell type, cells from a given tissue, and cells from a given organ, the correlation reflecting a change in the nature and amount of nucleic acids present in the classes, the method comprising the steps of:

a) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

b) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining a difference between the classes;

c) evaluating the correlation between the cells of the classes; and

d) preparing a representation of the correlation.

9. The method described in claim 8 wherein the determining of the presence and quantitation of the fragments described in step b) is carried out by a process comprising the steps of:

i) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (a) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (b) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

ii) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (a) the subsequences of the pairs of restriction endonucleases (b) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (c) the quantitation of the fragment corresponding to the pair and the length; and

iii) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (b) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

10. The method described in claim 8 wherein the determining of the presence of the fragments and the quantitation of the fragments, described in step b) is carried out by a process comprising the steps of:

i) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

ii) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

iii) generating output signals for each amplicon, each output signal characterizing (a) the subsequences of the pairs of primers, (b) the length, and (c) the quantitation; and

iv) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (b) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

11. The method described in claim 8 wherein the correlation in step d) is related to a set of orthonormal eigenvectors, the elements of the basis set upon which the eigenvectors are constructed reflecting particular biochemical or physiological pathways correlated between the cells of the two classes, each eigenvector having an eigenvalue that is an integer greater than zero, the coefficients of the basis set elements in each eigenvector whose eigenvalue is less than or equal to a particular integer that is an upper limit of the eigenvalues used reflecting the contribution of the corresponding pathway to the biochemical or physiological differences correlated between the cells of the first class and the cells of the second class.

12. The method described in claim 8 wherein the representation is a cluster diagram or a dendrogram, includes a tree structure reflecting the relatedness of the pathways involved in the biochemical or physiological response to a difference between cells of the two classes, wherein a correlation matrix provides a distance determination wherein the distance reflects the amplitude of a difference vector that is a difference between two vectors each of which reflects information obtained for the response of one of the two classes to the difference, and wherein the branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

13. The method described in claim 8 wherein the cells in at least one class are cancer cells.

14. The method described in claim 8 wherein the cells in at least one class have been contacted with a putative pharmaceutical agent, and the method comprises the steps of:

a) treating the cells of at least one class with an amount of the agent sufficient to effect a change in the state of those cells or with an amount of the agent less than or equal to a predetermined upper limit of dosing concentration;

b) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

c) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining an effect of the agent;

d) evaluating the correlation between the effect of the agent on the cells of the first class and the effect of the agent on the cells of another class; and

e) preparing a representation of the correlation.

15. A display means displaying a representation of the extent of relatedness between at least two classes of cells, wherein the cells in each class are chosen from the group consisting of cells of a given cell type, cells from a given tissue, and cells from a given organ, the extent of relateuness reflecting, in the nucleic acids of the classes of cells, similarities or differences in the presence of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence, a nucleotide length separating the first and second subsequences of the pair and a quantitation of the extent to which each pair having the determined length is in the classes of cells.

16. The display means described in claim 15 wherein the extent of relatedness is related to a distance wherein the distance reflects the amplitude of a difference vector that is a difference between a first vector which reflects information derived from the quantitation for each subsequence pair obtained for the first class and a second vector which reflects information derived from the quantitation for each subsequence pair obtained for the second class, wherein different elements of each vector relate to data obtained using different pairs.

17. The display means described in claim 15 wherein the representation includes a tree structure reflecting the relatedness between any two classes, and wherein the branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

18. The display means described in claim 15 wherein the extent of relatedness is obtained by a process comprising the steps of

a) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

b) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present; and

c) determining the extent of relatedness reflecting similarities or differences in the presence and quantitation of the fragments among the classes

19. The display means described in claim 18 wherein the determining of the presence and quantitation of the fragments described in step b) is carried out by a process comprising the steps of:

i) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (a) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (b) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

ii) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (a) the subsequences of the pairs of restriction endonucleases (b) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (c) the quantitation of the fragment corresponding to the pair and the length; and

iii) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (b) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

20. The display means described in claim 18 wherein the determining of the presence of the fragments and the quantitation of the fragments, described in step b) is carried out by a process comprising the steps of:

i) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

ii) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to

the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

iii) generating output signals for each amplicon, each output signal characterizing (a) the subsequences of the pairs of primers, (b) the length, and (c) the quantitation; and

iv) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (a) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

21. The display means described in claim 15 wherein the cells in at least one class are cancer cells.

22. The display means described in claim 15 wherein the cells in at least one class have been contacted with a putative pharmaceutical agent.

23. A display means displaying a representation of the correlation between a plurality of classes of cells, wherein the cells in each class are chosen from the group consisting of cells of a given cell type, cells from a given tissue, and cells from a given organ, the correlation reflecting, in the nucleic acids of the classes of cells, differences in the presence of a pair of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence and the nucleotide length separating the first and second subsequences of the pair, and a quantitation of the extent to which each pair having the determined length is present in the cells, between the classes.

24. The display means described in claim 23 wherein the correlation is related to a set of orthonormal eigenvectors, the elements of the basis set upon which the eigenvectors are constructed reflecting particular biochemical or physiological pathways correlated between the cells of the two classes, each eigenvector having an eigenvalue that is an integer greater than zero, the coefficients of the basis set elements in each eigenvector whose eigenvalue is less than a particular integer that is chosen to be an upper limit of the eigenvalues reflecting the contribution of the corresponding pathway to the biochemical or physiological differences correlated between the cells of the first class and the cells of the second class.

25. The display means described in claim 23 wherein the representation is a cluster diagram or a dendrogram and includes a tree structure reflecting the relatedness of the pathways involved in the biochemical or physiological difference between cells of the two classes, wherein a correlation matrix provides a distance determination wherein the distance reflects the amplitude of a difference vector that is a difference between two vectors each of which reflects information obtained for the difference between the classes, and wherein the branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

26. The display means described in claim 23 wherein the correlation is obtained by a method comprising the steps of

a) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

b)) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining a difference between classes;

c evaluating the correlation between the cells of one class and the cells of a second class based on the difference between them; and

d)) preparing a representation of the correlation.

27. The display means described in claim 23 wherein the determining of the presence and quantitation of the fragments described in step b) is carried out by a process comprising the steps of:

i) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (a) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (b) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

ii) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (a) the subsequences of the pairs of restriction endonucleases (b) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (c) the quantitation of the fragment corresponding to the pair and the length; and

iii) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (b) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

28. The display means described in claim 23 wherein the determining of the presence of the fragments and the quantitation of the fragments, described in step b) is carried out by a process comprising the steps of:

i) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

ii) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

iii) generating output signals for each amplicon, each output signal characterizing (a) the subsequences of the pairs of primers, (b) the length, and (c) the quantitation; and

iv) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (a) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

29. The display means described in claim 23 wherein the cells in at least one class are cancer cells.

30. The display means described in claim 23 wherein the cells in at least one class have been contacted with a putative pharmaceutical agent, and the correlation is obtained by method comprising the steps of

a) contacting the cells of at least one class with an amount of the agent sufficient to effect a change in the state of those cells or with an amount of the agent less than or equal to a predetermined upper limit of dosing concentration;

b) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

c) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining an effect of the agent;

d) evaluating the correlation between the effect of the agent between the cells of at least one class contacted with the agent and the cells of another class; and

e) preparing a representation of the correlation.

31. A representation of the extent of relatedness between at least two classes of cells, wherein the cells in each class are chosen from the group consisting of cells of a given cell type, cells from a given tissue, and cells from a given organ, the extent of relatedness reflecting, in the nucleic acids of the classes of cells, similarities or differences in the presence of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence, a nucleotide length separating the first and second subsequences of the pair and a quantitation of the extent to which each pair having the determined length is in the classes of cells.

32. The representation described in claim 31 wherein the extent of relatedness is related to a distance wherein the distance reflects the amplitude of a difference vector that is a difference between a first vector which reflects information derived from the quantitation for each subsequence pair obtained for the first class and a second vector which reflects information derived from the quantitation for each subsequence pair obtained for the second class, wherein different elements of each vector relate to data obtained using different pairs.

33. The representation described in claim 31 wherein the representation includes a tree structure reflecting the relatedness between any two classes, and wherein the branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

34. The representation described in claim 31 wherein the extent of relatedness is obtained by a process comprising the steps of

a) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

b) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present; and

c) determining the extent of relatedness reflecting similarities or differences in the presence and quantitation of the fragments among the cl sses

35. The representation described in claim 34 wherein the determining of the presence and quantitation of the fragments described in step b) is carried out by a process comprising the steps of:

i) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (a) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (b) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

ii) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (a) the subsequences of the pairs of restriction endonucleases (b) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (c) the quantitation of the fragment corresponding to the pair and the length; and

iii) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (b) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

36. The representation described in claim 34 wherein the determining of the presence of the fragments and the quantitation of the fragments, described in step b) is carried out by a process comprising the steps of:

i) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

ii) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to

the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

iii) generating output signals for each amplicon, each output signal characterizing (a) the subsequences of the pairs of primers, (b) the length, and (c) the quantitation; and

iv) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (a) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals.

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

37. The representation described in claim 31 wherein the cells in at least one class are cancer cells.

38. The representation described in claim 31 wherein the cells in a class have been contacted with a putative pharmaceutical agent.

38. A representation of the correlation between a plurality of classes of cells, wherein the cells in each class are chosen from the group consisting of cells of a given cell type, cells from a given tissue, and cells from a given organ, the correlation reflecting, in the nucleic acids of the classes of cells, differences in the presence of a pair of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence and the nucleotide length separating the first and second subsequences of the pair, and a quantitation of the extent to which each pair having the determined length is present in the cells, between the classes.

39. The representation described in claim 38 wherein the correlation is related to a set of orthonormal eigenvectors, the elements of the basis set upon which the eigenvectors are constructed reflecting particular biochemical or physiological pathways correlated between the cells of the two classes, each eigenvector having an eigenvalue that is an integer greater than zero, the coefficients of the basis set elements in each eigenvector whose eigenvalue is less than a particular integer that is chosen to be an upper limit of the eigenvalues reflecting the contribution of the corresponding pathway to the biochemical or physiological differences correlated between the cells of the first class and the cells of the second class.

40. The representation described in claim 38 wherein the representation is a cluster diagram or a dendrogram and includes a tree structure reflecting the relatedness of the pathways involved in the biochemical or physiological differences between cells of the two classes, wherein a correlation matrix provides a distance determination wherein the distance reflects the amplitude of a difference vector that is a difference between two vectors each of which reflects information obtained from one of the classes, and wherein the branches of the tree structure reflect the difference vectors and the branches are ramified from nodes.

41. The representation described in claim 38 wherein the correlation is obtained by a method comprising the steps of

a) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

b) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining a difference between classes;

c) evaluating the correlation between the cells of one class and the cells of a second class based on the difference between them; and

d) preparing a representation of the correlation.

42. The representation described in claim 41 wherein the determining of the presence and quantitation of the fragments described in step b) is carried out by a process comprising the steps of:

i) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (a) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (b) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

ii) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (a) the subsequences of the pairs of restriction endonucleases (b) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (c) the quantitation of the fragment corresponding to the pair and the length; and

iii) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (b) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

43. The representation described in claim 41 wherein the determining of the presence of the fragments and the quantitation of the fragments, described in step c) is carried out by a process comprising the steps of:

i) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

ii) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

iii) generating output signals for each amplicon, each output signal characterizing (a) the subsequences of the pairs of primers, (b) the length, and (c) the quantitation; and

iv) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (a) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (a) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

44. The representation described in claim 38 wherein the cells in at least one class are cancer cells.

45. The representation described in claim 38 wherein the cells in at least one class have been contacted with a putative pharmaceutical agent, and the correlation is obtained by a method comprising the steps of

a) contacting the cells of at least one class with an amount of the agent sufficient to effect a change in the state of those cells or with an amount of the agent less than or equal to a predetermined upper limit of dosing concentration;

b) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

c) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining an effect of the agent;

d) evaluating the correlation between the effect of the agent between the cells of at least one class contacted with the agent and the cells of another class; and

e) preparing a representation of the correlation.

46. A method for generating a geometrical representation between a plurality of classes of cells wherein the cells in each class are chosen from the group consisting of cells of a given cell type, cells from a given tissue, and cells from a given organ, the representation reflecting a change in the nature and amount of nucleic acids present in the classes, the method comprising the steps of:

a) in the nucleic acid of each class of cells, assessing the presence and amount of a nucleic acid fragment thereby defining a difference between the classes;

b) carrying out a geometrical analysis based on the differences between the cells of the classes; and

c) preparing a representation of the results of the analysis.

47. The method described in claim 46 wherein the geometrical representation is a result obtained by a principal component analysis or a principal factor analysis.

48. The method described in claim 46 wherein assessing the presence and amount of a nucleic acid fragment described in step a) is carried out by a process comprising the steps of:

i) probing the nucleic acid of each class with a set of oligonucleotide probes specific for the fragment; and

ii) determining the extent to which each probe binds the nucleic acid;

thereby providing an assessment of the presence and amount of the nucleic acid fragment in the class.

49. The method described in claim 46 wherein assessing the presence and amount of a nucleic acid fragment described in step a) is carried out by a process comprising the steps of::

i) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence; and

ii) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining the difference between the classes.

50. The method described in claim 49 wherein assessing the presence and quantity of a nucleic acid fragment described in step ii) is carried out by a process comprising the steps of:

(a) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (1) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (2) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

(b) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (1) the subsequences of the pairs of restriction endonucleases (2) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (3) the quantitation of the fragment corresponding to the pair and the length; and

(c) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (1) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (2) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals.

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

51. The method described in claim 49 wherein assessing the presence and quantity of a nucleic acid fragment described in step ii) is carried out by a process comprising the steps of:

(a) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

(b) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

(c) generating output signals for each amplicon, each output signal characterizing (1) the subsequences of the pairs of primers, (2) the length, and (3) the quantitation; and

(d) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (1) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (2) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals.

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

52. The method described in claim 46 wherein the results of the geometrical analysis are chosen from the group consisting of eigenvalues, eigenvectors, and principal factors.

53. The method described in claim 46 wherein the results of analysis in step c) are related to a set of orthonormal eigenvectors, the elements of the basis set upon which the eigenvectors are constructed reflecting particular biochemical, physiological or pharmacological components correlated between the cells of the two classes, each eigenvector having an eigenvalue, the coefficients of the basis set elements in each eigenvector reflecting the contribution of the corresponding biochemical, physiological or pharmacological components to the differences between the cells of the first class and the cells of the second class.

54. The method described in claim 46 wherein the cells in at least one class are cancer cells.

55. The method described in claim 46 wherein the cells in at least one class are contacted with a putative pharmaceutical agent, and the method comprises the steps of:

a) treating the cells of at least one class with an amount of the agent sufficient to effect a change in the state of those cells or with an amount of the agent less than or equal to a predetermined upper limit of dosing concentration;

b) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

c) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining an effect of the agent;

d) conducting a principal component analysis between the effect of the agent on the cells of the first class and the cells of another class; and

e) preparing a representation of the results of the analysis.

56. A display means displaying a geometrical representation between a plurality of classes of cells wherein the cells in each class are chosen from the group consisting of cells of a given cell type, cells from a given tissue, and cells from a given organ, the principal component analysis reflecting a change in the nature and amount of nucleic acids present in the classes, wherein the representation is obtained by a method comprising the steps of:

a) in the nucleic acid of each class of cells, assessing the presence and amount of a nucleic acid fragment thereby defining a difference between the classes;

b) carrying out a principal component analysis based on the differences between the cells of the first class and the cells of the second class; and

c) preparing the representation of the results of the analysis.

57. The display means described in claim 56 wherein the geometrical representation is a result obtained by a principal component analysis or a principal factor analysis.

58. The display means described in claim 56 wherein assessing the presence and amount of a nucleic acid fragment described in step a) comprises the steps of:

i) probing the nucleic acid of each class with a set of oligonucleotide probes specific for the fragment; and

ii) determining the extent to which each probe binds the nucleic acid;

thereby providing an assessment of the presence and amount of the nucleic acid fragment in the class.

59. The display means described in claim 56 wherein assessing the presence and amount of a nucleic acid fragment described in step a) is carried out by a process comprising the steps of:

i) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence; and

ii) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining the difference between the classes.

60. The display means described in claim 59 wherein determining the presence and quantity of a nucleic acid fragment described in step ii) is carried out by a process comprising the steps of:

(a) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (1) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (2) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

(b) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (1) the subsequences of the pairs of restriction endonucleases (2) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (3) the quantitation of the fragment corresponding to the pair and the length; and

(c) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (1) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (2) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

61. The display means described in claim 59 wherein assessing the presence and quantity of a nucleic acid fragment described in step ii) is carried out by a process comprising the steps of:

(a) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

(b) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

(c) generating output signals for each amplicon, each output signal characterizing (1) the subsequences of the pairs of primers, (2) the length, and (3) the quantitation; and

(d) optionally searching a nucleotide sequence database to determine sequences that are predicted to oduce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (1) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (2) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals,

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

62. The display means described in claim 56 wherein the results of the analysis are chosen from the group consisting of eigenvalues, eigenvectors, and principal factors.

63. The display means described in claim 56 wherein the results of the analysis in step c) are related to a set of orthonormal eigenvectors, the elements of the basis set upon which the eigenvectors are constructed reflecting particular biochemical, physiological or pharmacological components correlated between the cells of the two classes, each eigenvector having an eigenvalue, the coefficients of the basis set elements in each eigenvector reflecting the contribution of the corresponding biochemical, physiological or pharmacological components to the differences between the cells of the first class and the cells of the second class.

64. The display means described in claim 56 wherein the cells in at least one class are cancer cells.

65. The display means described in claim 56 wherein the cells in at least one class have been contacted with a putative pharmaceutical agent, and the representation is obtained by a method comprising the steps of:

a) treating the cells of at least one class with an amount of the agent sufficient to effect a change in the state of those cells or with an amount of the agent less than or equal to a predetermined upper limit of dosing concentration;

b) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

c) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining an effect of the agent;

d) conducting a principal component analysis between the effect of the agent on the cells of the first class and the cells of another class; and

e) preparing the representation of the results of the analysis.

66. A geometrical representation between a plurality of classes of cells wherein the cells in each class are chosen from the group consisting of c ''s of a given cell type, cells from a given tissue, and cells from a given organ, the principal component analysis reflecting a change in the nature and amount of nucleic acids present in the classes, the representation obtained by a method comprising the steps of:

a) in the nucleic acid of each class of cells, assessing the presence and amount of a nucleic acid fragment thereby defining a difference between the classes:

b) carrying out a principal component analysis based on the differences between the cells of the first class and the cells of the second class; and

c) preparing the representation of the results of the analysis.

67. The representation described in claim 66 wherein the geometrical representation is a result obtained by a principal component analysis or a principal factor analysis.

68. The representation described in claim 66 wherein assessing the presence and amount of a nucleic acid fragment described in step a) comprises the steps of:

i) probing the nucleic acid of each class with a set of oligonucleotide probes specific for the fragment; and

ii) determining the extent to which each probe binds the nucleic acid;

thereby providing an assessment of the presence and amount of the nucleic acid fragment in the class.

69. The representation described in claim 66 wherein assessing the presence and amount of a nucleic acid fragment described in step a) is carried out by a process comprising the steps of:

i) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence; and

ii) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining the difference between the classes.

70. The representation described in claim 69 wherein determining the presence and quantity of a nucleic acid fragment described in step ii) is carried out by a process comprising the steps of:

(a) digesting samples of the nucleic acid from the cells of each class with a plurality of specific pairs of restriction endonucleases, each sample being treated by one pair, one nuclease of the pair targeting the first subsequence and the second nuclease of the pair targeting the second subsequence, each digestion providing specific restriction fragments, hybridizing double stranded adapter DNA molecules to the fragments, each adapter DNA molecule comprising (1) a shorter strand having no 5' terminal phosphate and consisting of a first and second portion, said first portion being at the 5' end and being complementary to the overhang produced by one of the restriction endonucleases of the pair, and (2) a longer strand having a 3' end complementary to the second portion of the shorter strand, and ligating the longer strands to the fragments to produce ligated fragments, wherein each ligated fragment is capable of generating an output signal;

(b) generating output signals from each ligated fragment for each of the pairs of restriction endonucleases, each output signal characterizing (1) the subsequences of the pairs of restriction endonucleases (2) the length between the two subsequences corresponding to the two restriction endonucleases employed in each pair of nucleases, and (3) the quantitation of the fragment corresponding to the pair and the length; and

(c) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (1) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (2) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals.

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains fragments having the specific subsequence pairs and the nucleotide length between the pairs.

71. The representation described in claim 69 wherein assessing the presence and quantity of a nucleic acid fragment described in step ii) is carried out by a process comprising the steps of:

(a) for each pair of nucleotide subsequences providing a pair of oligonucleotide primers, consisting of a first primer and a second primer, wherein the first primer is complementary to the first subsequence and the second primer is complementary to the second subsequence;

(b) amplifying the nucleotide sequence between the first subsequence and the second subsequence using the oligonucleotide primers to prime the amplification, providing an amplicon characterized by the subsequence pair, a length between the two subsequences corresponding to the two primers employed in each pair and a quantitation of the extent to which each amplicon is present; and

(c) generating output signals for each amplicon, each output signal characterizing (1) the subsequences of the pairs of primers, (2) the length, and (3) the quantitation; and

(d) optionally searching a nucleotide sequence database to determine sequences that are predicted to produce or the absence of any sequences that are predicted to produce the one or more output signals produced by the nucleic acid from the cells of each class, the database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the cells of each class, a sequence from the database being predicted to produce the one or more output signals when the sequence from the database has both (1) the same length between occurrences of target nucleotide subsequences as is represented by the one or more output signals, and (2) the same target nucleotide subsequence as are represented by said one or more output signals, or target nucleotide subsequences that are members of the same stets of target nucleotide subsequences represented by the one or more output signals.

thereby providing a quantitative measure of the extent to which the nucleic acid present in the cells in each class contains the specific subsequence pairs and the nucleotide length between the pairs.

72. The representation described in claim 66 wherein the results of the analysis are chosen from the group consisting of eigenvalues, eigenvectors, and principal factors.

73. The representation described in claim 66 wherein the results of the analysis in step c) are related to a set of orthonormal eigenvectors, the elements of the basis set upon which the eigenvectors are constructed reflecting particular biochemical, physiological or pharmacological components correlated between the cells of the two classes, each eigenvector having an eigenvalue, the coefficients of the basis set elements in each eigenvector reflecting the contribution of the corresponding biochemical, physiological or pharmacological components to the differences between the cells of the first class and the cells of the second class.

74. The representation described in claim 66 wherein the cells in at least one class are cancer cells.

75. The representation described in claim 66 wherein the cells in at least one class have been contacted with a putative pharmaceutical agent, and the representation is obtained by a method comprising the steps of:

a) treating the cells of at least one class with an amount of the agent sufficient to effect a change in the state of those cells or with an amount of the agent less than or equal to a predetermined upper limit of dosing concentration;

b) defining a plurality of pairs of nucleotide subsequences, each pair consisting of a first subsequence and a second subsequence;

c) in the nucleic acid of each class of cells determining the presence of a fragment with the first subsequence at one end and the second subsequence at another end and having a length separated by the first and second subsequences, and a quantitation of the extent to which each fragment is present, thereby defining an effect of the agent;

d) conducting a principal component analysis between the effect of the agent on the cells of the first class and the cells of another class; and

e) preparing the representation of the results of the analysis.

76. A method for classifying a plurality of classes of cells or components thereof hierarchically comprising the steps of

a) measuring relative differences in the quantity of a nucleic acid present in each class of cells to provide measurements of differential nucleic acid display;

b) converting the measurements into distances between the classes of cells in a vector space; and

c) preparing a hierarchical classification amongst the classes based on the vector distances.

77. The method of claim 76 wherein the classification is performed on classes of cells, wherein the cells in a class may be cells of a given cell type, cells from a given tissue, and cells from a given organ, cells exhibiting a particular pathological state, or cells which have been contacted with a putative pharmaceutical agent.

78. The method of claim 76 wherein the classification is performed on a component of the cells in the classes, wherein the component comprises a gene, a nucleic acid, or a fragment thereof.

79. The method of claim 76 wherein the measuring is carried out by a procedure chosen from the group consisting of differential display of nucleic acid fragments, probing for the presence of a nucleic acid using an oligonucleotide probe, sequences obtained from expressed sequence tags (ESTs), assessing restriction fragment length polymorphisms, and assessing amplification fragment length polymorphisms

80. The method of claim 76 wherein the preparation of the hierarchical classification is carried out by a procedure chosen from the group consisting of principal component analysis of a correlation matrix, principal factor analysis of a correlation matrix, principal component analysis of a centered inner product matrix, and principal factor analysis of a centered inner product matrix.

81. The method of claim 80 further comprising the step of obtaining a distance metric between the classes from a reduced dimensionality geometrical representation.

82. A display means displaying the results of the classification obtained by a method described in any one of claims 76-81.

83. A method for representing a plurality of classes of cells or components thereof geometrically comprising the steps of

a) measuring relative differences in the quantity of a nucleic acid present in each class of cells to provide measurements of differential nucleic acid display; and

b) preparing a geometrical representation amongst the classes based on the measurement of the differential display.

84. The method of claim 83 wherein the classification is performed on classes of cells, wherein the cells in a class may be cells of a given cell type, cells from a given tissue, and cells from a given organ, cells exhibiting a particular pathological state, or cells which have been contacted with a putative pharmaceutical agent.

85. The method of claim 83 wherein the classification is performed on a component of the cells in the classes, wherein the component comprises a gene, a nucleic acid, or a fragment thereof.

86. The method of claim 83 wherein the measuring is carried out by a procedure chosen from the group consisting of differential display of nucleic acid fragments, probing for the presence of a nucleic acid using an oligonucleotide probe, sequences obtained from expressed sequence tags (ESTs), assessing restriction fragment length polymorphisms, and assessing amplification fragment length polymorphisms

87. The method of claim 83 wherein the preparation of the hierarchical classification is carried out by a procedure chosen from the group consisting of principal component analysis of a correlation matrix, principal factor analysis of a correlation matrix, principal component analysis of a centered inner product matrix, and principal factor analysis of a centered inner product matrix.

88. The method of claim 87 further comprising the step of obtaining a distance metric between the classes from a reduced dimensionality geometrical representation.

89. A display means displaying the results of the geometrical representation obtained by a method described in any one of claims 83-88.

90. A method of presenting the hierarchical relatedness of two or more members of a population, the method comprising:

providing a data set of each member in the population:

generating a hierarchical classification of said data set; and

displaying said classification, thereby presenting the hierarchical relatedness of the members of the population.

91. The method of claim 90, wherein said population is a population of cells.

92. The method of claim 90, wherein said population is a population of nucleic acid sequences.

93. The method of claim 90, wherein said population is a population of polypeptide sequences.

94. The method of claim 90, wherein said hierarchical classification of any two or more members of the population is calculated using a distance method in combination with an algorithm.

95. The method of claim 94, wherein said distance method is a Pearson correlation distance, Euclidean distance, Manhattan distance, Mahalanobis distance, a pairwise Pearson distance, or a Spearman distance.

96. The method of claim 95, wherein said algorithm is single linkage, average linkage, or complete linkage.

97. The method of claim 90, wherein said data set is the product of an analysis of said members of the population that is selected from the group consisting of differential display, serial analysis of gene expression , expression tagged sequence analysis, restriction fragment length polymorphism, amplified fragment length polymorphism, or Northern blot hybridization analysis.

98. A method of presenting the geometrical relatedness of two or more members of a population, the method comprising:

providing a data set of each member in the population;

generating a geometrical classification of said data set; and

displaying said classification, thereby presenting the geometrical relatedness of the members of the population.

99. The method of claim 98, wherein said population is a population of cells.

100. The method of claim 98, wherein said population is a population of nucleic acid sequences.

101. The method of claim 98, wherein said population is a population of polypeptide sequences.

102. The method of claim 98, wherein said geometrical classification is generated by analyzing a matrix using an algorithm.

103. The method of claim 102, wherein said matrix includes a correlation matrix.

104. The method of claim 103, wherein said correlation matrix includes a Pearson correlation matrix, a Spearman correlation matrix, or a pairwise Pearson correlation matrix.

105. The method of claim 102, wherein said matrix includes a centered inner product distance matrix.

106. The method of claim 105, wherein the inner product distance matrix is determined using a distance calculated by hierarchical classification analysis.

107. The method of claim 102, wherein said algorithm includes principal component analysis.

108. The method of claim 102, wherein said algorithm includes principal factor analysis.

109. The method of claim 107, wherein said algorithm includes principal factor analysis.

110. The method of claim 102, wherein said geometrical classification is further analyzed using hierarchical classification.

111. The method of claim 90, wherein said population includes 5, 10, 25, 50, 100, 1000, 10,000, 100,000 or more members.

112. The method of claim 98, wherein said population includes 5, 10, 25, 50, 100, 1000, 10,000, 100,000 or more members.

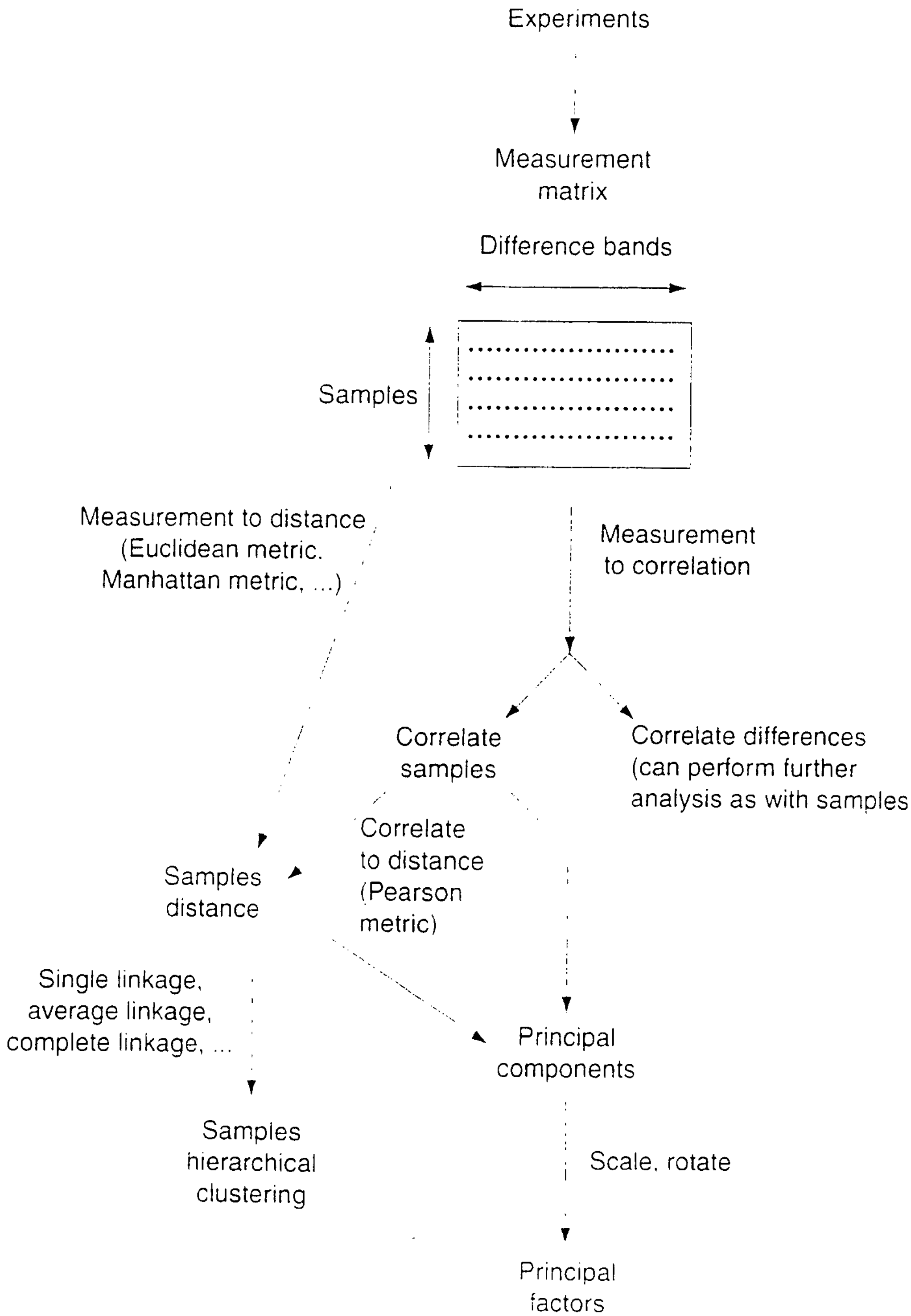

Experiments
Measurement matrix
Difference bands
Samples
Measurement to distance (Euclidean metric. Manhattan metric, ...)
Measurement to correlation
Correlate samples
Correlate differences (can perform further analysis as with samples)
Correlate to distance (Pearson metric)
Samples distance
Single linkage, average linkage, complete linkage, ...
Principal components
Samples hierarchical clustering
Scale, rotate
Principal factors

Fig. 1

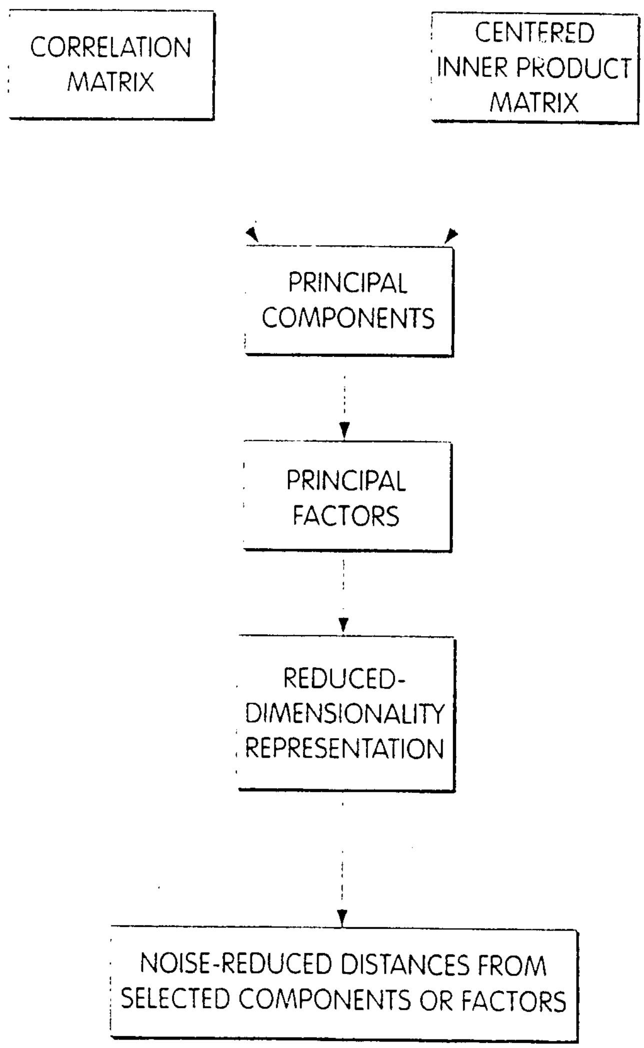

CORRELATION MATRIX
CENTERED INNER PRODUCT MATRIX
PRINCIPAL COMPONENTS
PRINCIPAL FACTORS
REDUCED-DIMENSIONALITY REPRESENTATION
NOISE-REDUCED DISTANCES FROM SELECTED COMPONENTS OR FACTORS

Fig. 2

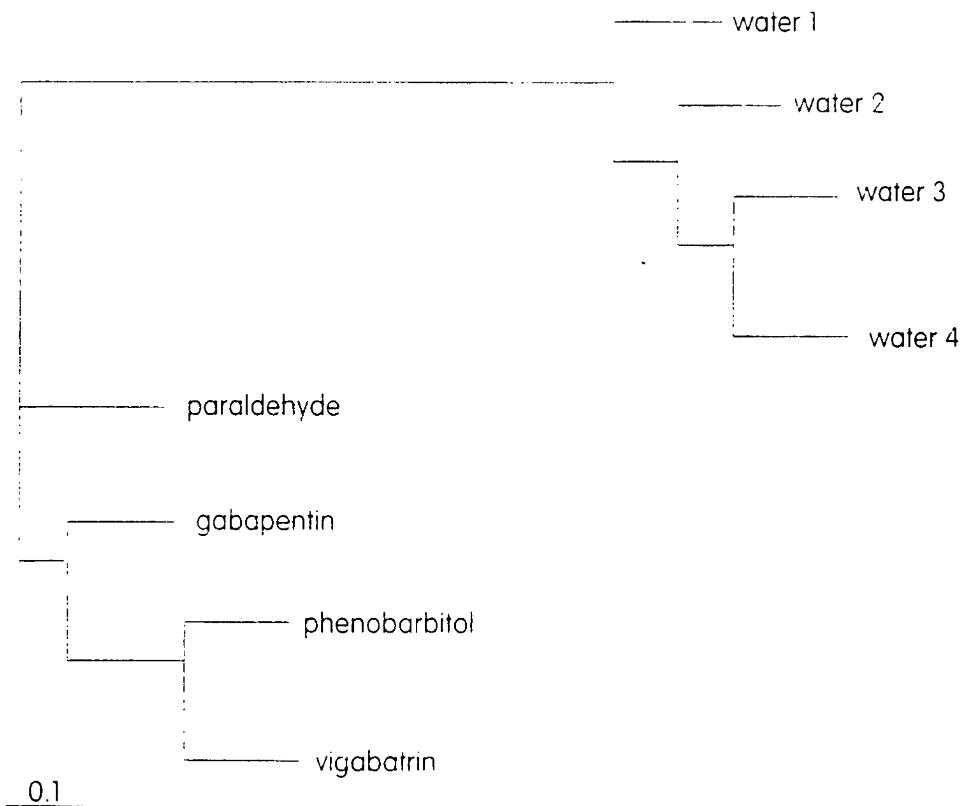

water 1
water 2
water 3
water 4
paraldehyde
gabapentin
phenobarbitol
vigabatrin
0.1

Fig. 3

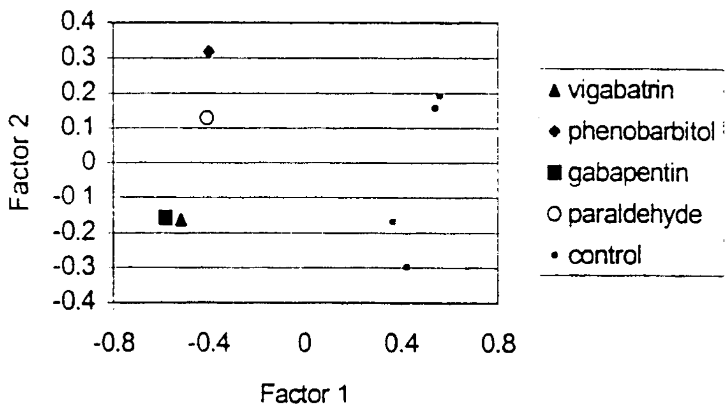

0.4
0.3
0.2
0.1
0
-0 1
-0.2
-0.3
-0.4
Factor 2
-0.8
-0.4
0
0.4
0.8
Factor 1
vigabatrin
phenobarbitol
gabapentin
paraldehyde
control

Fig. 4

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 99/21525

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12Q1/68 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12Q G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 97 15690 A (CURAGEN CORP) 1 May 1997 (1997-05-01) the whole document ---	1-112
X	GUILFOYLE R A ET AL: "Ligation-mediated PCR amplification of specific fragments from class-II restriction endonuclease total digest" NUCLEIC ACIDS RESEARCH, XP002076198 the whole document ---	1-112
	-/--	

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
9 February 2000	28/02/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Müller, F

2

Form PCT/ISA/210 (second sheet) (July 1992)

Intern. Application No
PCT/US 99/21525

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KATO K: "DESCRIPTION OF THE ENTIRE MRNA POPULATION BY A 3' END CDNA FRAGMENT GENERATED BY CLASS IIS RESTRICTION ENZYMES" NUCLEIC ACIDS RESEARCH,GB,OXFORD UNIVERSITY PRESS, SURREY, vol. 23, no. 18, 1 September 1995 (1995-09-01), pages 3685-3690, XP002008304 ISSN: 0305-1048 the whole document	1-112
X	WO 97 22720 A (BEATTIE KENNETH LOREN) 26 June 1997 (1997-06-26) the whole document	1-112
X	WO 97 29211 A (US HEALTH ;WEINSTEIN JOHN N (US); BOULAMWINI JOHN (US)) 14 August 1997 (1997-08-14) the whole document	1-112
X	WO 97 13877 A (LYNX THERAPEUTICS INC ;MARTIN DAVID W (US)) 17 April 1997 (1997-04-17) the whole document	1-112
A	US 5 508 169 A (DEUGAU KENNETH V ET AL) 16 April 1996 (1996-04-16) the whole document	
P,X	SHIMKETS R.A. ET AL.,: "Gene expression analysis by transcript profiling coupled to a gene database query" NATURE BIOTECHNOLOGY, vol. 17, - August 1999 (1999-08) pages 798-803, XP002130008 cited in the application the whole document	1-112

2

INTERNATIONAL SEARCH REPORT

Information on patent family members

Intern. Application No
PCT/US 99/21525

Patent document cited in search report		Publication date	Patent family member(s)			Publication date
WO 9715690	A	01-05-1997	US	5871697	A	16-02-1999
			US	59726[illegible]3	A	26-10-1999
			AU	7476396	A	15-05-1997
			EP	0866877	A	30-09-1999
WO 9722720	A	26-06-1997	AU	1687597	A	14-07-1997
WO 9729211	A	14-08-1997	AU	2264197	A	28-08-1997
WO 9713877	A	17-04-1997	AU	712929	B	18-11-1999
			AU	4277896	A	06-05-1996
			AU	6102096	A	30-12-1996
			AU	7717596	A	30-04-1997
			CN	1193357	A	16-09-1998
			CZ	9700866	A	17-09-1997
			CZ	9703926	A	17-06-1998
			EP	0793718	A	10-09-1997
			EP	0832287	A	01-04-1998
			EP	0931165	A	28-07-1999
			FI	971473	A	04-06-1997
			HU	9900910	A	28-07-1999
			JP	11507528	T	06-07-1999
			JP	10507357	T	21-07-1998
			NO	971644	A	02-06-1997
			NO	975744	A	05-02-1998
			PL	324000	A	27-04-1998
			WO	9641011	A	19-12-1996
US 5508169	A	16-04-1996	US	5858656	A	12-01-1999
			CA	2036946	A	07-10-1991